

# GAUSSIAN ORACLE INEQUALITIES FOR STRUCTURED SELECTION IN NON-PARAMETRIC PARTIAL LIKELIHOOD

JELENA BRADIC AND RUI SONG

**ABSTRACT.** To better understand the interplay of censoring and sparsity we develop finite sample properties of nonparametric Cox proportional hazard's model. Due to high impact of sequencing data, carrying genetic information of each individual, we work with over-parametrized problem and propose general class of group penalties suitable for sparse structured variable selection and estimation. Novel non-asymptotic sandwich bounds for the partial likelihood are developed. We establish how they extend notion of local asymptotic normality (LAN) of Le Cam's. Such non-asymptotic LAN principles are further extended to high dimensional spaces where  $p \gg n$ . Finite sample prediction properties of penalized estimator in non-parametric Cox proportional hazards model, under suitable censoring conditions, agree with those of penalized estimator in linear models.

## 1. INTRODUCTION

Sparse linear models have emerged as powerful framework to deal with high dimensional problems where the size of parameter space  $p$  is much larger than the sample size  $n$ , i.e. for the case of  $p \gg n$ . Its applications range from machine learning to signal processing and provide a necessary tool to analyze new, high-throughput data emerging from various scientific fields through online auction data or functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) signals or new techniques of mRNA sequencing. The approach of  $l_1$  regularization has become a popular tool to efficiently address such problems by assuming sparsity in parameter space (Tibshirani, 1996; Fan and Li, 2001; Meinshausen and Yu, 2009; Lv and Fan, 2009; Zhou, 2010). If number of sparse elements is not our only constraint different forms of group or hierarchical regularization have been proposed to handle structured information (see for example Yuan and Lin (2006); Zhao et al. (2009)). In the general sparsity setting, finite sample predictive properties of penalized least squares procedures have been well studied and understood. Finite sample oracle inequalities with and without group structure have been obtained even for overparametrized problems where  $p \gg n$  (see for example Bickel et al. (2009); Meier et al. (2009); Ravikumar et al. (2009); Kolar et al. (2011); Jenatton et al. (2011); Lounici et al. (2011); Raskutti et al. (2012)). While a large body of work has focused on parametric and non-parametric linear models, censored high dimensional data have been left fairly unexplored when  $p \gg n$ . Moreover, driven by the known affect that censoring rate has on sample reduction, we were interested in discovering finite sample properties of censored models and their possible dependence on censoring rate.

Estimation of high-dimensional proportional hazards ratio can be classified into two categories: univariate and regularized multivariate methods and their analysis into asymptotic and non-asymptotic kind. Univariate or marginal methods in predicting genetic abundances that are linked to survival

---

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF STATISTICS, COLORADO STATE UNIVERSITY

*E-mail address:* jbradic@math.ucsd.edu; song@stat.colostate.edu .

have been used for decades without much understanding of the underlying possibility for severe overestimation. [Zhao and Li \(2012\)](#) outlined basic assumptions needed for univariate hazards models to have model selection consistency. As expected conditions are quite restrictive and significantly impair the applicability of such methodology. Our focus here is non-asymptotic analysis of multivariate regularized methods for right censored data where we aim for an estimator that can accurately predict nonparametric covariate effects. Popular approach is to look at  $l_1$ -norm regularized log likelihood (see [Tibshirani \(1996\)](#); [Fan and Li \(2001\)](#); [Yuan and Lin \(2006\)](#)). Multivariate methods have been analyzed recently in [Bradic et al. \(2011\)](#) where the emphasis was on asymptotical model selection properties of non-convex regularized partial likelihood. Non-asymptotic oracle properties have been developed for different additive hazards models ([Gaïffas and Guillaou, 2012](#); [Kong and Nan, 2012](#); [Lemler, 2012](#)) where weighted  $l_1$  norm penalty was employed to encourage sparsity and likelihood function was approximated with  $l_2$  type of criterion. To the best of our knowledge non-asymptotic, finite sample prediction properties of multivariate methods for Cox like partial likelihoods have not been studied when the dimensionality of the covariates is ultra-high. They pose significant challenges as the partial likelihood doesn't obey  $l_2$  criterion and strict convexity cannot be guaranteed in the whole high dimensional space.

To that end, we consider bivariate data  $\{(X_i, T_i) : i = 1, \dots, n\}$ , which form an i.i.d sample from the population  $(\mathbf{X}, \mathbf{T})$  and where  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  is a column vector of  $p$ -covariate processes for the  $i$ -th individual. We are interested in the cases where not all survival times  $(T_i)_{i=1}^n$  are fully observable and where independent right censoring scheme is assumed (censoring times  $(C_i)_{i=1}^n$  are conditionally on  $(\mathbf{X}_i)_{i=1}^n$  independent of survival times  $(T_i)_{i=1}^n$ ). Hence we work with i.i.d sample of the form

$$\{(\mathbf{X}_i, Z_i, \delta_i) : i = 1, \dots, n\},$$

where  $Z_i = \min(T_i, C_i)$  and  $\delta_i = \mathbf{1}\{T_i \leq C_i\}$  are event times and censoring indicator respectively. Note that we do not allow for time-dependent covariates, and will work under the boundedness assumptions of the form  $\mathbf{X} = [a, b]^p$  (for some constants  $a, b$ ). Conditional hazard function of  $T$  given  $\mathbf{X} = x$  is denoted as  $\lambda(t|x)$  and is defined as instantaneous rate of failure at time  $t$  given a particular value  $x$  of covariate  $\mathbf{X}$ . We are interested in non-parametric hazards model, where the effect of covariate on the intensity process can be described as :

$$(1) \quad \lambda(t|x) = \lambda_0(t)g(x),$$

for a baseline hazard function  $\lambda_0(t)$  as conditional hazard function of  $T$  given  $x = 0$  (common covariate effect on the survival time) and a relative risk function  $g : \mathcal{R}^p \rightarrow \mathcal{R}_+$ .

We propose a general methodology for analyzing such complex models, with identifying local neighborhood where strict convexity can be guaranteed and then analyzing predictive properties in the whole space by chaining through local neighborhood. Contributions of our paper are three fold.

(1) We establish new general and sparse oracle inequalities for high dimensional regularized right censored non-parametric Cox model (1). New type of empirical functional norms are defined and used to describe the predictive performance of the proposed estimator. Classical GOI and SOI results, i.e. those obtained for least squares regression problems, become only local properties in models with complex log-likelihood structure. Locality is increased with the increase of dimensionality  $p$  of the problem.

(2) We develop new two-stage technique and extend Le Cam's ([Le Cam, 1960](#)) Local Asymptotic Normality (LAN) to high dimensional spaces through something we call, local non-asymptotic approach. We discuss this approach through Cox model where we prove sandwich bounds for

log partial-likelihood, bounding it from below and from above with quadratic random processes. First layer is composed of processes that are not necessarily i.i.d and not necessarily generated by Gaussian linear models. The second layer is developed with the help of the first and is composed of processes that are truly Gaussian in nature. This new technique provides tools for connecting GOI and SOI results obtained for this model and for least squares regression models. Interestingly, we show that this connection holds in local elliptical neighborhood of  $\beta^*$  and doesn't hold outside of such sets if the dimensionality is high. By localizing our estimator we were able to extend this new idea to problems where  $p \gg n$  but  $\log p \leq n$ .

(3) We show explicitly how new non-asymptotic predictive performance of penalized estimator is connected to the censoring rate of the model. Asymptotic theory is unable to retrieve such connections due to disappearance of censoring rate with the increase of the sample size. To that end results presented in this paper are novel and unique. They provide direct relationship between finite sample risk properties and observed number of uncensored data. This relationship is linear in local neighborhood where local non-asymptotic normality holds. Over the whole space, the relationship is more complicated and is directly embedded in the proposed risk functions.

To handle high dimensionality of the problem we introduce general group penalty where we treat differently convex and non-convex types of regularization. Closer relations to our work are primary regarding sparse oracle inequalities for regularized Gaussian type log-likelihoods (see for example [van de Geer \(2008\)](#); [Bickel et al. \(2009\)](#); [Bunea et al. \(2009\)](#); [Jenatton et al. \(2011\)](#); [Gaïffas and Guilloux \(2012\)](#)) where the quadratic structure of the loss function eases of technical difficulties. Localization technique we developed is complementary to the recent work of [Spokoiny \(2012\)](#) but is based on different non-asymptotic scheme that greatly relaxes assumptions and simplifies the structure of the bounding processes presented in the latter paper. Additionally, we make connections to the likelihoods that become truly Gaussian in the full observational study, and further extend our method to the problems with high dimensional structure.

The paper is organized as follows. In this Section we introduce notation and model setup and define new empirical functional norms adapted to the setting of censored data. In Section 2 we define local neighborhood and propose non-asymptotic extension of LAN property with two types of sandwich bounds. In Section 3 and 4 we develop slow and fast rate convergent novel oracle inequalities. We use these inequalities to localize our estimator to small neighborhood where we prove gaussian type prediction properties. In that way we show that even in high dimensional spaces, when censoring is not severe, gaussian type of prediction properties hold in sparse censored Cox proportional hazards models. They match Gaussian case, where we observe all data points. In Section 5 we propose non-convex regularized estimator and further extend previous finite sample oracle inequalities. Section 6 is left for particular examples .

**1.1. Model Formulation.** Let us consider high dimensional, non-parametric Cox hazards regression model, where the hazard rate function  $\lambda(t|x)$  takes the form  $\lambda(t|x) = \lambda_0(t)g(x)$  with  $g(x) \in \mathcal{G}_n$ ,

$$(2) \quad \mathcal{G}_n = \{g : [e^a, e^b]^p \rightarrow \mathcal{R} : \log g(x) = f(x), f \in \mathcal{F}_n\},$$

where  $\mathcal{F}_n$  is the collection of functions  $f$  on  $[a, b]^p$  with the additive structure  $f(x) = f_1(x_1) + \dots + f_p(x_p)$ . Goal is to estimate unknown functions  $f_j(x)$  using best linear combination of a finite dictionary of candidate functions  $\Psi_1(x), \dots, \Psi_d(x)$ . Constraints on the candidate functions  $\Psi$  are that they are known a priori and bounded above with a constant  $C$ . Note that we can always center the data so that they are mean zero. In this way we want to design a statistical learning procedure that adapts to the unknown functions  $f_j$  for which we only assume it belongs to a linear space spanned by our dictionary (see for example [Bunea et al. \(2007\)](#), [Rigollet \(2012\)](#)). By standard

notation let us define  $N_i(t) = 1\{Z_i \leq t, \delta_i = 1\}$ ,  $\bar{N}(t) = n^{-1} \sum_{i=1}^n N_i(t)$ , and  $Y_i(t) = 1\{Z_i \geq t\}$ . We may then write the likelihood as

$$\begin{aligned} \mathcal{P}_\tau(\mathbf{b}) &= \prod_{t \in \mathcal{F}} \prod_{i=1}^n d\Lambda(t)^{\Delta N_i(t)} [1 - d\bar{\Lambda}(t)]^{1 - \Delta \bar{N}(t)} \\ (3) \quad &= \prod_{t \in \mathcal{F}} \prod_{i=1}^n [\Lambda_0(t) \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\}]^{\Delta N_i(t)} [1 - \Lambda_0(t) \mathcal{S}_n^{(0)}(\mathbf{b}, t)]^{1 - \Delta \bar{N}(t)}, \end{aligned}$$

where we take the following vector notation  $\mathbf{b} = [\mathbf{b}_1^T | \mathbf{b}_2^T | \cdots | \mathbf{b}_p^T]$  with  $\mathbf{b}_j = (b_{j1}, \dots, b_{jd})^T$  and  $\boldsymbol{\Psi}(\mathbf{X}_i) = [\boldsymbol{\Psi}(X_{i1})^T | \cdots | \boldsymbol{\Psi}(X_{ip})^T]$  with  $\boldsymbol{\Psi}(X_{ij}) = (\Psi_1(X_{ij}), \dots, \Psi_d(X_{ij}))^T$  and for  $l = 0, 1, 2$  with  $\otimes$  denoting outer product

$$\mathcal{S}_n^{(l)}(\mathbf{b}, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \boldsymbol{\Psi}^{\otimes l}(\mathbf{X}_i) \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\}.$$

The Nelson-Aalen estimator of the integrated underlying hazard function  $\hat{\Lambda}_0(t, \mathbf{b}) = \int_0^t d\bar{N}(s) \times \mathcal{S}_n^{(0)-1}(\mathbf{b}, s)$  becomes Breslow type of estimator which allows us to write the normalized log partial likelihood as

$$(4) \quad \mathcal{L}_n(\mathbf{b}, \tau) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i) dN_i(t) - \int_0^\tau \log \mathcal{S}_n^{(0)}(\mathbf{b}, t) d\bar{N}(t),$$

where and hereafter  $\tau$  is fixed study end time.

**1.2. Family of Folded Group Penalties.** Let us define the empirical risk function  $\mathcal{R}_n(\mathbf{b})$  as negative of log partial likelihood

$$(5) \quad \mathcal{R}_n(\mathbf{b}) = -\mathcal{L}_n(\mathbf{b}),$$

where from now on for every fixed time window  $\tau$ ,  $\mathcal{L}_n(\mathbf{b}, \tau) = \mathcal{L}_n(\mathbf{b})$ . Since the maximization of the log-likelihood is an overparametrized problem for  $p \gg n$  we introduce a penalty structure  $P(\mathbf{b})$  and hence search for the estimator  $\hat{\boldsymbol{\beta}}$  of the unknown parameter, as a solution of the following problem

$$(6) \quad \min_{\mathbf{b} \in \mathcal{R}^{p \times d}} \{\mathcal{R}_n(\mathbf{b}) + \lambda_n P(\mathbf{b})\},$$

where the penalty function  $P(\mathbf{b})$  is defined as

$$(7) \quad P(\mathbf{b}) = \sum_{j=1}^p \text{df}_j \cdot \rho(\|\mathbf{b}_j\|_{\gamma_j}) = \sum_{j=1}^p \text{df}_j \cdot \rho\left(\left\{\sum_{k=1}^d |b_{jk}|^{\gamma_j}\right\}^{1/\gamma_j}\right).$$

Parameters  $\text{df}_j$  are the group scaling corresponding to the degrees of freedom of each group. The typical choice of  $\text{df}_j$ , for convex functions  $\rho$ , is  $d^{1/\gamma_j^*}$ , for  $\gamma_j^*$  being a Hölder conjugate of  $\gamma_j$ , to ensure that the penalty term is of the order of the number of parameters  $\text{df}_j$ .

The additive model of the hazard function  $\lambda$  in (2) induces the natural grouping structure over the parameter space  $\{b_{jk}\}_{j,k=1}^{p,d}$  at the functional level. When regularizing by Lasso penalty, sparsity is induced by treating each variable individually, and existing relationships and structures between the variables (spatial, hierarchical or related to the physics of the problem) are merely disregarded (see Zhao et al. (2009); Jenatton et al. (2011)). Hence, a penalty that encourages sparsity across functional groups is more suitable.

A family of folded group penalty functions (FGP) (7) takes into account existing structure in the parameter space and encourages sparsity in the additive model. In its full generality the family of FGP penalties, can induce a wide variety of grouping structures in the coefficients  $\{b_{jk}\}_{j,k=1}^{p,d}$ :  $\rho$  determines how the groups relate to one another, while  $\{L_{\gamma_j}\}_{j=1}^p$  norms dictate the relationship of the coefficients among each group  $j$ . In this way, while the number of groups is minimized whereas the variables in each group are selected as a whole block. This follows easily from the well known properties of bridge estimators for  $\gamma_j \geq 1$ , where  $\gamma_j$  determines how close together the size of coefficients in each selected group are kept.

In that respect for specific functions  $\rho$  and parameters  $\gamma_j$  FGP (7) reduces to a number of already proposed group penalties. For example, for  $\rho = L_1$ , and any  $\gamma_j$  it reduces to the CAP family  $\lambda \sum_{j=1}^p \|f_j\|_{\gamma_j}$  of Zhao et al. (2009); for  $\rho = L_1, \gamma_j = 2$  it becomes the group Lasso penalty  $\lambda \sum_{j=1}^p \|f_j\|_2$  of Yuan and Lin (2006); Obozinski et al. (2010); Lounici et al. (2011); for  $\rho = L_1$  and  $\gamma_j = \infty$  it reduces to block  $l_1/l_\infty$  penalty  $\lambda \sum_{j=1}^p \|f_j\|_\infty$  of Zhang et al. (2008) and Negahban and Wainwright (2011). Moreover, the problem can be reparametrized to include a scaling in the penalty function of the following form  $\rho(\|\mathbf{b}_j^T \mathbf{R}_j\|_{\gamma_j})$  with  $\mathbf{R}_j = (\mathbf{R}_1(X_j) | \dots | \mathbf{R}_d(X_j))^T$ ,  $\mathbf{R}_k(X_j) = (R_k(X_{i1}), \dots, R_k(X_{nj}))^T$ , for some weighting functions  $\{R_k\}_{k=1}^d : \mathcal{R} \rightarrow \mathcal{R}$ . Similar scaling was used in structured group Lasso in van de Geer (2011) where it was combined with plain Lasso estimator to obtain sparse solutions. Moreover, as smoothing is necessary for the B-spline basis expansion Meier et al. (2009), the penalty in (7) can be easily adapted to include such scaling as  $\rho(\|\mathbf{R}_j \mathbf{b}_j\|_{\gamma_j} + \sqrt{\mathbf{b}_j^T \mathbf{M}_j \mathbf{b}_j})$  where the  $d \times d$  smoothing matrix  $\{\mathbf{M}_j\}_{kl} = \int \Psi_k''(x_j) \Psi_l''(x_j) dx_j$ . All the results in the paper can be extended to hold for both such extensions (see Section 6), but for simplicity of notation we will be working with the penalty function as defined in (7).

**1.3. Notation.** Let  $\mathbf{b}$  stand for the  $p \times d$  vector of parameters  $\{b_{jk}\}_{j,i=1}^{p,n}$  such that  $f_{\mathbf{b}}(x) = \sum_{j=1}^p \sum_{k=1}^d b_{jk} \Psi_k(x)$ . Let  $f^0(x)$  now denote the  $p$ -dimensional additive approximation of the unknown function  $f(x)$  with the following structure

$$(8) \quad f^0(x) = f_{\beta}(x) = \sum_{j=1}^p \sum_{k=1}^d \beta_{jk} \Psi_k(x_j),$$

where the unknown parameter vector  $\beta = [\beta_1^T | \dots | \beta_p^T]$  and  $\beta_j = (\beta_{j1}, \dots, \beta_{jd})^T$ . With this we have moved away from the additive model (2) to the fully nonparametric model with  $\lambda(t|x) = \lambda_0(t) \exp\{f(x)\}$ . To avoid curse of dimensionality (commonly recognized in nonparametric problems), the risk function in (5) serves only as a proxy to the fully nonparametric Cox negative partial likelihood. Naturally, by allowing  $p \gg n$  we are able to achieve better approximation to the maximum of the fully nonparametric partial log likelihood. With this notation we have

$$f^0(\mathbf{X}_i) = f_{\beta}(\mathbf{X}_i) = \sum_{j=1}^p \sum_{k=1}^d \beta_{jk} \Psi_k(X_{ij}) = \beta^T \Psi(\mathbf{X}_i),$$

for a  $p * d$  dimensional vector  $\Psi(\mathbf{X}_i)$  as described in the description of equation (9). In that sense, let  $\beta^*$  denote the sparse alternative to  $\beta$ , which corresponds to  $f_{\beta^*}(x)$  being sparse equivalent of the additive function  $f^0(x)$  i.e. sparse additive approximation of the unknown function  $f(x)$  (see Meier et al. (2009)). This sparse approximation,  $f_{\beta^*}(x)$ , is equal to  $\sum_{j=1}^s f_j^*(x) =$

$\sum_{j=1}^s \sum_{k=1}^d \beta_{jk}^* \Psi_k(x_j)$ . i.e.  $f_{\beta^*}(\mathbf{X}_i) = \beta^{*T} \Psi(\mathbf{X}_i)$  with  $\beta^* = [\beta_1^{*T} |\beta_2^{*T}| \cdots |\beta_s^{*T}| \mathbf{0}^T \cdots \mathbf{0}^T]$ ,  $\beta_j^* = (\beta_{j1}, \dots, \beta_{jd})^T$ , and  $\mathbf{0} = (0, \dots, 0)^T \in \mathcal{R}^d$ .

Let us define an empirical functional norm  $\|\cdot\|_{n, \mathbf{b}^*}$  of all functions  $f_{\mathbf{b}} : R^{p*d} \rightarrow R$  for any  $\mathbf{b} \in R^p$  and a fixed  $\mathbf{b}^* \in R^{p*d}$  as

$$\|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 = \frac{1}{n} \sum_{i=1}^n \int_0^\tau Y_i(t) \omega_i(\mathbf{b}^*, t) f_{\mathbf{b}}^2(\mathbf{X}_i) d\bar{N}(t) - \left[ \frac{1}{n} \sum_{i=1}^n \int_0^\tau Y_i(t) \omega_i(\mathbf{b}^*, t) f_{\mathbf{b}}(\mathbf{X}_i) d\bar{N}(t) \right]^2,$$

for nonnegative weight process

$$\omega_i(\mathbf{b}^*, t) = \exp\{f_{\mathbf{b}^*}(\mathbf{X}_i)\} / S_n^{(0)}(\mathbf{b}^*, t).$$

The intuition behind the introduction of this norm is given in Section 2.1. It will be crucial in analyzing fine non-asymptotic properties of penalized estimator  $\hat{\beta}$  and is connected to the curvature of the log-likelihood process. Because no two counting processes  $N_i(t)$  jump at the same time, the following holds

$$(9) \quad \|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 = \frac{1}{n} \sum_{i=1}^n \int_0^\tau Y_i(t) \omega_i(\mathbf{b}^*, t) (f_{\mathbf{b}}(\mathbf{X}_i) - \bar{f}_{\mathbf{b}}^*(t))^2 d\bar{N}(t),$$

where  $\bar{f}_{\mathbf{b}}^*(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \omega_i(\mathbf{b}^*, t) f_{\mathbf{b}}(\mathbf{X}_i)$  can be understood as a process of empirical weighted averages of  $f_{\mathbf{b}}$ . Notice that, if Condition 1 is satisfied (see Section 2.1) the introduced empirical norm exhibits all the properties of proper norm. It is nonnegative,  $\|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 = 0$  for every  $\mathbf{b}^*$  if and only if  $\mathbf{b} = \mathbf{0}$  (see Condition 1(iii) in Section 2.1) and satisfies triangular inequality  $\|f_{\mathbf{b}_1} - f_{\mathbf{b}_2}\|_{n, \mathbf{b}^*} \leq \|f_{\mathbf{b}_1}\|_{n, \mathbf{b}^*} + \|f_{\mathbf{b}_2}\|_{n, \mathbf{b}^*}$  for every  $\mathbf{b}_1, \mathbf{b}_2$  and any fixed  $\mathbf{b}^*$  (by simple algebraic manipulations).

Let us denote with  $t_1 < \dots < t_N$  ordered failure times and with  $\mathcal{R}_q = \{i \in \{1, \dots, n\} : Z_i \geq t_q\}$  risk sets. Then we can define a  $l_2$  empirical functional norm  $\|\cdot\|_n^2$  for the case of censored data as follows

$$(10) \quad \|f_{\mathbf{b}}\|_n^2 = \frac{1}{n} \sum_{i=1}^n 1\{i \in \mathbf{I}\} f_{\mathbf{b}}^2(\mathbf{X}_i),$$

where  $\mathbf{I} = \{i \in \{1, \dots, n\} : i \in \cup_{q=1}^N \mathcal{R}_q\}$ . If we are willing to assume that  $\mathbf{I} = \{1, \dots, n\}$  (as it seems natural to consider only patients that belong to at least one risk set), the previous definition (10) matches classical functional  $l_2$  norm of the form  $\frac{1}{n} \sum_{i=1}^n f_{\mathbf{b}}^2(\mathbf{X}_i)$ .

## 2. LOCAL NON-ASYMPTOTIC BOUNDS

This chapter expands the ideas of Le'Cam's local asymptotic normality (LAN) in two directions. First, it develops a non-asymptotic equivalent of LAN techniques where instead of proving convergence to a single gaussian event, we show finite sample gaussian sandwich bounds. Finite sample techniques are important in modern high throughput data where dimensionality of the parameter space prohibits us to analyze methods asymptotically. In high dimensional problems, it is not clear if local perturbations of  $1/\sqrt{n}$  rate are the optimal ones as our intuition suggests that the rates of convergence might be of the order of  $\sqrt{\log p/n}$  when  $p \gg n$  but  $\log p \leq n$  and trivial extension of classical asymptotic approach fails due to vectors and matrices of exploding size. Hence, we develop non-asymptotic extension of it, that allows us to define finite sample sandwich bounds on the likelihood process.

We show that dimensionality  $p$  plays crucial role even in local bounds and that local neighborhood of Le'Cam type are inherently dimensionality dependent in cases where model structure is complex

and where  $p \gg n$  (see Section 3 for more details). Furthermore, we take these local bounds and combine them with sparse oracle inequalities to show that penalized estimator can have predictive properties close to the oracle estimator of gaussian linear models, i.e. simple linear model with Gaussian errors. In this sense, we extend LAN idea to overparamterized problems with  $p \gg n$  (see Section 4 for more details).

**2.1. Local Non-Asymptotic Quadraticity.** In this section we will show important quadratic interpretation of the risk function  $\mathcal{R}_n(\mathbf{b})$ . We will show work for the partial log-likelihood (4) but note that this representation of  $\mathcal{R}_n(\mathbf{b})$  works for many complex log and log partial likelihood problems and all the ideas can be easily generated to specific cases. Note that without loss of generality  $\mathcal{R}_n(\mathbf{b})$  can be written as  $\mathcal{R}_n(\mathbf{b}) = -\mathcal{L}_n(\mathbf{b}) + \mathcal{L}_n(\beta^*) - \mathcal{L}_n(\beta^*)$ . By Taylor expansion around  $\beta^*$  we have that there exists a  $c \in (0, 1)$  and  $\mathbf{b}^* = c\mathbf{b} + (1-c)\beta^*$  such that

$$\mathcal{R}_n(\mathbf{b}) = -(\mathbf{b} - \beta^*)^T \{\nabla \mathcal{L}_n(\beta^*)\} - \frac{1}{2} (\mathbf{b} - \beta^*)^T \{\nabla^2 \mathcal{L}_n(\mathbf{b}^*)\} (\mathbf{b} - \beta^*) - \mathcal{L}_n(\beta^*).$$

Let us introduce the following notation  $\mathbf{E}_n(\mathbf{b}, t) = S_n^{(1)}(\mathbf{b}, t)/S_n^{(0)}(\mathbf{b}, t)$  and  $\mathbf{V}_n(\mathbf{b}, t) = S_n^{(2)}(\mathbf{b}, t)/S_n^{(0)}(\mathbf{b}, t) - (S_n^{(1)}(\mathbf{b}, t)/S_n^{(0)}(\mathbf{b}, t))^{\otimes 2}$ . With this notation at hand, the score vector and the hessian of the log partial likelihood have the following representations respectively

$$\begin{aligned} -\{\nabla \mathcal{L}_n(\mathbf{b})\} &= n^{-1} \sum_{i=1}^n \int_0^\tau (\mathbf{E}_n(\mathbf{b}, t) - \Psi(\mathbf{X}_i)) dN_i(t), \\ -\{\nabla^2 \mathcal{L}_n(\mathbf{b})\} &= n^{-1} \sum_{i=1}^n \int_0^\tau \mathbf{V}_n(\mathbf{b}, t) dN_i(t). \end{aligned}$$

With the notation introduced in the previous section and by simple algebraic manipulations, the following quadratic representation of the hessian matrix holds:

$$-\mathbf{b}^T \{\nabla^2 \mathcal{L}_n(\mathbf{b}^*)\} \mathbf{b} = \|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2.$$

Together with previous Taylor expansion, we have that the empirical risk function decomposes as follows:

$$\begin{aligned} \mathcal{R}_n(\mathbf{b}) &= n^{-1} \sum_{i=1}^n \int_0^\tau (\mathbf{b} - \beta^*)^T (\mathbf{E}_n(\beta^*, t) - \Psi(\mathbf{X}_i)) dN_i(t) \\ &\quad - \frac{1}{2n} \sum_{i=1}^n \int_0^\tau (\mathbf{b} - \beta^*)^T \mathbf{V}_n(\mathbf{b}^*, t) (\mathbf{b} - \beta^*) dN_i(t) - \mathcal{L}_n(\beta^*). \end{aligned}$$

That is, for every  $\mathbf{b}$  there exists a  $c \in (0, 1)$  and  $\mathbf{b}^* = c\mathbf{b} + (1-c)\beta^*$  such that  $\mathcal{R}_n(\mathbf{b})$  admits the following quadratic representation:

$$(11) \quad \mathcal{R}_n(\mathbf{b}) = -(\mathbf{b} - \beta^*)^T \mathbf{h}_n(\beta^*) + \frac{1}{2} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 - \mathcal{L}_n(\beta^*),$$

where we have the concatenated score vector  $\mathbf{h}_n(\beta^*) = \{\nabla \mathcal{L}_n(\beta^*)\} \in \mathbb{R}^{p \times d}$  defined as:

$$(12) \quad \mathbf{h}_n(\beta^*) = n^{-1} \sum_{i=1}^n \int_0^\tau (\mathbf{E}_n(\beta^*, t) - \Psi(\mathbf{X}_i)) dN_i(t),$$

with,  $\mathbf{h}_n(\beta^*) = [\mathbf{h}_{n,1}^T(\beta^*) | \mathbf{h}_{n,2}^T(\beta^*) | \dots | \mathbf{h}_{n,p}^T(\beta^*)]$ , where each  $\mathbf{h}_{n,j}^T(\beta^*)$  is a  $d$ -dimensional vector  $\mathbf{h}_{n,j}^T(\beta^*) = (\{\mathbf{h}_n(\beta^*)\}_{j1}, \{\mathbf{h}_n(\beta^*)\}_{j2}, \dots, \{\mathbf{h}_n(\beta^*)\}_{jd})^T$ .



To examine the properties of quadratic representation of the risk function  $\mathcal{R}_n(\mathbf{b})$  and its connection to a quadratic form that doesn't involve parameter  $\mathbf{b}^*$  we define the process

$$(13) \quad \mathcal{G}_{n,\eta}(\mathbf{b}) = -(\mathbf{b} - \beta^*)^T \mathbf{h}_n(\beta^*) + \frac{1}{2}\eta \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\beta^*}^2 - \mathcal{L}_n(\beta^*),$$

which is a natural extension to the right hand side of equation (11). We look at how well can  $\mathcal{G}_{n,\eta}(\mathbf{b})$  approximate the risk function  $\mathcal{R}_n(\mathbf{b})$  for any vector  $\mathbf{b}$ . This approximation also tells us how the hessian matrix  $\mathbf{V}_n(\mathbf{b}^*)$  is related to  $\mathbf{V}_n(\beta^*)$ , which is not a problem for linear regression with quadratic empirical risk  $\mathcal{R}_n(\mathbf{b})$  as  $\mathbf{V}_n(\mathbf{b}^*) = \mathbf{V}_n(\beta^*)$  for all  $\mathbf{b}^*$ . The key of this approximation is in the following proposition.

**PROPOSITION 1.** *Let  $a_{\mathbf{v}}$  for any vector  $\mathbf{v}$  be  $a_{\mathbf{v}} = \max_{1 \leq i, q \leq n} |\mathbf{v}^T [\Psi(\mathbf{X}_i) - \Psi(\mathbf{X}_q)]|$ . Then the following sandwich bound holds almost surely for every vector  $\mathbf{b}$  and corresponding vector  $\mathbf{b}^* = c\mathbf{b} + (1-c)\beta^*$ ,*

$$(14) \quad e^{-2a_{\mathbf{b}-\beta^*}} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\beta^*}^2 \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 \leq e^{2a_{\mathbf{b}-\beta^*}} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\beta^*}^2.$$

*Proof of Proposition 1.* To see that the equation (14) is correct, we adopt the following reasoning. First, note that  $\|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2$  is equal to

$$n^{-1} \int_0^\tau \frac{\sum_{i,q=1}^n w_i w_q (a_i - a_q)^{\otimes 2} e^{(1-c)a_i - \bar{c}} e^{(1-c)a_q - \bar{c}}}{\sum_{i,q=1}^n 2w_i w_q e^{(1-c)a_i - \bar{c}} e^{(1-c)a_q - \bar{c}}} d\bar{N}(t),$$

with  $a_i = (\mathbf{b} - \beta^*)^T (\Psi(\mathbf{X}_i) - \mathbf{E}_n(\beta^*, t))$  and  $w_i = Y_i(t) \exp\{\beta^{*T} \Psi(\mathbf{X}_i)\}$  and  $\bar{c} = (1-c)(\max_i a_i + \min_i a_i)/2$ . If we let  $\eta = a_{\mathbf{b}-\beta^*}$  we can see that  $\max_i |(1-c)a_i - \bar{c}| \leq \eta/2$ . Using this notation we can see that from  $e^{(1-c)a_i - \bar{c}} \geq e^{-\eta/2}$  and  $e^{(1-c)a_i - \bar{c}} \leq e^{\eta/2}$  we have

$$\begin{aligned} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 &\geq \exp\{-2\eta\} n^{-1} \int_0^\tau \frac{\sum_{i,q=1}^n w_i w_q (a_i - a_q)^{\otimes 2}}{\sum_{i,q=1}^n 2w_i w_q} d\bar{N}(t) \\ &= \exp\{-2\eta\} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\beta^*}^2. \end{aligned}$$

Upper bound follows the same reasoning and is therefore omitted.  $\square$

From this proposition it follows that the process  $\mathcal{R}_n(\mathbf{b})$  is almost surely upper bounded by the process  $\mathcal{G}_{n,\eta}$  for  $\log \eta = 2a_{\mathbf{b}-\beta^*}$  and lower bounded by the process  $\mathcal{G}_{n,\eta}$  for  $\log \eta = -2a_{\mathbf{b}-\beta^*}$ . Original stochastic approximation problem is now sandwiched between two quadratic stochastic approximation problems of the following kind:

$$\mathcal{G}_{n,e^{-2a_{\mathbf{b}-\beta^*}}}(\mathbf{b}) + \lambda_n P(\mathbf{b}) \leq \mathcal{R}_n(\mathbf{b}) + \lambda_n P(\mathbf{b}) \leq \mathcal{G}_{n,e^{2a_{\mathbf{b}-\beta^*}}}(\mathbf{b}) + \lambda_n P(\mathbf{b}).$$

Moreover, studying properties of  $\mathcal{R}_n(\hat{\beta})$  can be translated to studying the properties of the two quadratic processes  $\mathcal{G}_{n,e^{-2a_{\mathbf{b}-\beta^*}}}(\hat{\beta})$  and  $\mathcal{G}_{n,e^{2a_{\mathbf{b}-\beta^*}}}(\hat{\beta})$ . The definition (13) indicates that the processes  $\mathcal{G}_{n,\eta}$  have a geometric structure similar to log-likelihood of Gaussian model. However, they are not necessarily equal in distribution to Gaussian process since the norm  $\|\cdot\|_{n,\beta^*}$  incorporates data which didn't come from Gaussian model. It turns out that their geometric structure is of more importance than their distributional qualities.



**2.2. Local Non-Asymptotic Normality.** Previously defined processes  $\mathcal{G}_{n,\eta}$  although having quadratic geometric structure are not Gaussian distributed. In case of simple linear model with Gaussian errors, they become equal to  $\mathcal{G}_{n,1}$  which with big probability follow Gaussian distribution. However, in case of Cox proportional hazards model, due to the complicated  $\eta$  term, they do not follow Gaussian distribution. In order to better understand their relationship to Gaussian processes and dimension  $p$ , we turn to relationship between introduced empirical functional norms  $\|\cdot\|_{n,\mathbf{b}^*}$  and  $\|\cdot\|_n$ .

With slight abuse of notation, note that the weight process  $\omega_i(\mathbf{b}, t)$  and empirical norm  $\|f_{\mathbf{b}}\|_n$  are tightly connected to the following weighting vector  $\omega_i(\mathbf{b})$  expressed through the sequence of risk sets  $\{\mathcal{R}_q\}_{q=1}^N$  as follows

$$(15) \quad \omega_i(\mathbf{b}) = \sum_{q=1}^N \frac{\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\} 1\{i \in \mathcal{R}_q\}}{\sum_{l \in \mathcal{R}_q} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}}.$$

As sum of conditional probabilities that the observation  $i$  had an event at time  $t_q$ , given that at least one event occurred at time  $t_q$ , we know they are necessarily nonnegative. This representation will be useful for the characterization of Gaussian bounds on the log partial likelihood process in the following Proposition.

**PROPOSITION 2.** *Let  $N$  represent the number of distinct events. Then, the uniform sandwich bound for the norm  $\|f_{\mathbf{b}}\|_{n,\mathbf{b}^*}$ , for every  $\mathbf{b}$  and corresponding  $\mathbf{b}^* = c\mathbf{b} + (1-c)\beta^*$ , with  $c \in (0, 1)$ , hold almost surely*

$$(16) \quad \underline{\omega} \|f_{\mathbf{b}}\|_n^2 \leq \|f_{\mathbf{b}}\|_{n,\mathbf{b}^*}^2 \leq N \|f_{\mathbf{b}}\|_n^2,$$

where  $\|f_{\mathbf{b}}\|_n^2$  denotes the censored  $l_2$  empirical norm as defined in (10) and where

$$\underline{\omega} = \min_{\substack{\mathbf{b} \in \mathcal{R}^{p^*d}, c \in (0,1) \\ i \in \{1, \dots, n\} : i \in \cup_{q=1}^N \mathcal{R}_q}} \omega_i(\beta^* + c(\mathbf{b} - \beta^*)),$$

with sequence of risk sets  $\{\mathcal{R}_q\}_{q=1}^N$  as defined in Section 1 and  $\omega_i(\mathbf{b})$  defined in (15).

*Proof of Proposition 2.* Let  $N$  stands for the cardinality of the set  $\{i = 1, \dots, n : N_i(\tau) = 1\}$ . Note that the weight process  $\omega_i(\mathbf{b}, t)$  satisfy the following normalization uniformly over  $\mathbf{b}$  and  $t$ ,

$$\frac{1}{n} \sum_{i=1}^n Y_i(t) \omega_i(\mathbf{b}, t) = 1.$$

Moreover, for each  $\mathbf{b}$  there exists at least one  $i \in \{1, \dots, n\}$  such that  $\omega_i(\mathbf{b}, t) > 0$  and that for all  $i$ , for which  $\exists t \in [0, \tau]$ ,  $Y_i(t) = 1$ , we have that  $\omega_i(\mathbf{b}, t) \leq n$ .<sup>1</sup> Let us denote with

$$\omega_i(\mathbf{b}) = \int_0^\tau Y_i(t) \omega_i(\mathbf{b}, t) d\bar{N}(t).$$

If  $t_1 < \dots < t_N$  are ordered failure times and  $\mathcal{R}_j = \{i \in \{1, \dots, n\} : Z_i \geq t_j\}$  is at risk set, then  $\omega_i(\mathbf{b})$  has the following representation:

$$\omega_i(\mathbf{b}) = \sum_{j=1}^N \frac{\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\} 1\{i \in \mathcal{R}_j\}}{\sum_{l \in \mathcal{R}_j} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}}.$$

---

<sup>1</sup> Assume that there exists at least one  $i$  such that  $\omega_i(\mathbf{b}) > n$  and  $Y_i(t) = 1$ . Then,  $1 < \frac{1}{n} Y_i(t) \omega_i(\mathbf{b}) + \frac{1}{n} \sum_{q=1, q \neq i}^n Y_q(t) \omega_q(\mathbf{b}) = \frac{1}{n} \sum_{q=1}^n Y_q(t) \omega_q(\mathbf{b}) = 1$

Firstly, note that  $\omega_i \geq 0$  and  $\omega_i > 0$  for  $i \in \mathbf{I} = \{i \in \{1, \dots, n\} : i \in \cup_{j=1}^N \mathcal{R}_j\}$ . Using previous notation we have

$$\|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 = \frac{1}{n} \sum_{i=1}^n f_{\mathbf{b}}^2(X_i) \omega_i(\mathbf{b}^*) - \left( \frac{1}{n} \sum_{i=1}^n f_{\mathbf{b}}(X_i) \omega_i(\mathbf{b}^*) \right)^2,$$

With this notation at hand we have that

$$\frac{1}{n} \sum_{i=1}^n \omega_i(\mathbf{b}^*) = \frac{1}{n} \sum_{j=1}^N \sum_{i \in \mathcal{R}_j} \frac{\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{\sum_{l \in \mathcal{R}_j} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}} = \frac{N}{n}.$$

and we are able to conclude  $N \geq \bar{\omega} = \max\{\omega_i(\mathbf{b}) : i \in \mathbf{I}, \mathbf{b} \in R^{p*d}\} \geq N/|\mathbf{I}| \geq \underline{\omega} = \min\{\omega_i(\mathbf{b}) : i \in \mathbf{I}, \mathbf{b} \in R^{p*d}\}$  for  $\mathbf{I} = \{i \in \{1, \dots, n\} : i \in \cup_{j=1}^N \mathcal{R}_j\}$ . Hence,

$$\|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 \leq \bar{\omega} \|f_{\mathbf{b}}\|_n^2 - \underline{\omega} \left( \frac{1}{n} \sum_{i \in \mathbf{I}} f_{\mathbf{b}}(X_i) \right)^2 \leq N \|f_{\mathbf{b}}\|_n^2.$$

To obtain the left hand side of (16) remember that from previous exposition we have

$$\|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 = \frac{1}{n} \sum_{i=1}^n \omega_i(\mathbf{b}_{\mathbf{b}}) (f_{\mathbf{b}}(\mathbf{X}_i) - \bar{f}_{\mathbf{b}}^*)^2$$

with  $\bar{f}_{\mathbf{b}}^* = \frac{1}{n} \sum_{i=1}^n \omega_i(\mathbf{b}_{\mathbf{b}}) f_{\mathbf{b}}(X_i)$ . Hence, by centering the data so that the sample mean is equal to zero, that is  $\frac{1}{n} \sum_{i \in \mathbf{I}} f_{\mathbf{b}}(\mathbf{X}_i) = 0$ , we have

$$\begin{aligned} \|f_{\mathbf{b}}\|_{n, \mathbf{b}^*}^2 &\geq \underline{\omega} \frac{1}{n} \sum_{i \in \mathbf{I}} (f_{\mathbf{b}}^2(\mathbf{X}_i) + \{\bar{f}_{\mathbf{b}}^*\}^2) + 2\underline{\omega} \bar{f}_{\mathbf{b}}^* \left( \frac{1}{n} \sum_{i \in \mathbf{I}} f_{\mathbf{b}}(\mathbf{X}_i) \right) \\ &\geq \underline{\omega} \frac{1}{n} \sum_{i \in \mathbf{I}} f_{\mathbf{b}}^2(\mathbf{X}_i) = \underline{\omega} \|f_{\mathbf{b}}\|_n^2. \end{aligned}$$

□

The lower bound in the previous proposition is always non-trivial as  $\underline{\omega} \geq 0$  but there is very little hope that it is significantly bounded away from zero for all choices of parameter  $\mathbf{b}$ , in the whole  $pd$  dimensional space. To that end, note that risk sets  $\mathcal{R}_j$  are naturally nested as follows  $\mathcal{R}_1 \supseteq \dots \supseteq \mathcal{R}_N$  with  $\cup_{j=1}^N \mathcal{R}_j = \mathcal{R}_1$  and  $\cap_{j=1}^N \mathcal{R}_j = \mathcal{R}_N$ . Then, for  $i \in \mathcal{R}_1 \setminus \mathcal{R}_2$

$$\omega_i(\mathbf{b}) = \frac{\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{\sum_{l \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}} \geq \frac{\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{|\mathcal{R}_1| \max_{l \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}}$$

Following similar reasoning we have for  $i \in \mathcal{R}_2 \setminus \mathcal{R}_3$ , that is  $i \in \mathcal{R}_1$ ,  $i \in \mathcal{R}_2$ ,  $i \notin \mathcal{R}_3$ ,

$$\omega_i(\mathbf{b}) \geq \frac{2 \exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{|\mathcal{R}_1| \max_{l \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}}$$

where we used  $\max_{l \in \mathcal{R}_2} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\} \leq \max_{l \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}$  and  $|\mathcal{R}_1| \geq |\mathcal{R}_2|$ . By applying similar reasoning we can see that for  $i \in \mathcal{R}_N$

$$\omega_i(\mathbf{b}) \geq \frac{N \exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{|\mathcal{R}_1| \max_{l \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_l)\}}.$$

Knowing  $\min_{i \in \mathbf{I}} \omega_i(\mathbf{b}) = \min \{ \min_{i \in \mathcal{R}_1 \setminus \mathcal{R}_2} \omega_i(\mathbf{b}), \min_{i \in \mathcal{R}_2 \setminus \mathcal{R}_3} \omega_i(\mathbf{b}) \cdots, \min_{i \in \mathcal{R}_N} \omega_i(\mathbf{b}) \}$  we are left to analyze relative ordering of the minima on the right hand side of previous equality. It is obvious that the minima on the right hand side are non-decreasing, hence concluding that

$$\min_{i \in \mathbf{I}} \omega_i(\mathbf{b}) = \min_{i \in \mathcal{R}_1 \setminus \mathcal{R}_2} \omega_i(\mathbf{b}) \geq \frac{\min_{i \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}{|\mathcal{R}_1| \max_{i \in \mathcal{R}_1} \exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}}.$$

It is common to assume that all  $n$  observations belong to set  $\mathcal{R}_1$ , making its cardinality of size  $n$ . From the last inequality it becomes clear that when the range of random variables  $\exp\{\mathbf{b}^T \Psi(\mathbf{X}_i)\}$  becomes smaller,  $\underline{\omega}$  is more bounded away from zero. Guaranteeing that would amount to requiring  $\sup_{\mathbf{b} \in \mathcal{R}^{pd}} \|\mathbf{f}_{\mathbf{b}}\|_{\infty} \leq c \log n$ , for some positive constant  $c$ . Obviously such a requirement in high dimensional spaces restricts the set of functions  $f$  significantly. On the other hand, requiring that the sparse approximation  $f_{\beta^*}$  has the property of  $\|f_{\beta^*}\|_{\infty} \leq c \log n$ , requires  $s \leq \log n$ , as constant  $c$  is bounded (all  $\omega_i(\beta^*)$  are bounded random variables). Note that the usual upper bound on the size of the sparsity set is  $n$  but due to the complex problem structure we have to pay the price in requiring much smaller sparsity. Similar conclusion was reached in the recent paper [Bradic et al. \(2011\)](#) where they discovered that situation of  $s = 25, p = 1000$  and  $n = 100$  is not sparse enough and even the oracle estimator fails empirically, compared to  $s = 4, p = 1000, n = 100$  where the oracle estimator was behaving regularly. In that sense, next Corollary 1 gives a complete picture on how the sparsity size constraints the problem at hand and shows that the interplay between true and effective dimensionality in survival models is different from the case of linear regression.

**COROLLARY 1.** *Let  $N$  represent the number of distinct events. Then, the uniform sandwich bound for the norm  $\|f_{\mathbf{b}}\|_{n, \beta^*}$*

$$(17) \quad \underline{\omega} \|f_{\mathbf{b}}\|_n^2 \leq \|f_{\mathbf{b}}\|_{n, \beta^*}^2 \leq N \|f_{\mathbf{b}}\|_n^2,$$

where  $\underline{\omega} = \min\{\omega_i(\beta^*) : i \in \{1, \dots, n\}, i \in \cup_{q=1}^N \mathcal{R}_q\}$ .

Propositions 1 and 2 show a non-asymptotic parallel to Le Cam's theory of local asymptotic normality (LAN) ([Le Cam, 1960](#)), which shows that a statistical model can be locally approximated by a gaussian model. Le Cam's approximation is asymptotic in nature. Our extension is non-asymptotic, as it bounds our model from above and from bellow with geometrically Gaussian like models, that is we identified lower and upper processes  $\mathcal{G}_{n, \eta_1}(\mathbf{b})$ ,  $\mathcal{G}_{n, \eta_2}(\mathbf{b})$  such that there exists a ball  $\mathbb{B}_{\beta^*}(r_n)$  such that

$$P(\mathcal{G}_{n, \eta_1}(\mathbf{b}) \leq \mathcal{R}_n(\mathbf{b}) \leq \mathcal{G}_{n, \eta_2}(\mathbf{b}), \forall \mathbf{b} \in \mathbb{B}_{\beta^*}(r_n)) = 1.$$

Independent work that discusses non-asymptotic extension of Le Cam's theory appeared in [Spokoiny \(2012\)](#). Our work for Cox model can be viewed as a non-trivial extension of their work on general log-likelihood structures with the following reasoning. Their bounding processes are defined for  $p \ll n$  and by shrinking and expanding of the equivalence of Hessian matrix  $\nabla^2 \mathcal{L}_n$ . This approach is suitable for cases where such a matrix doesn't depend on parameter space. In other cases, such as the situation in our setup, it is not obvious how to define lower and upper bounding processes with this approach. Furthermore, by adapting steps of condition (ED<sub>1</sub>) and Theorem 3.1 of their approach to the Cox model, the bound achieved is not as tight as the approximation (16).<sup>2</sup>

<sup>2</sup> Parameter  $\omega(\mathbf{r})$  in condition (ED<sub>1</sub>) of [Spokoiny \(2012\)](#), would have to be chosen strictly positive, due to dependence of score vector on the parameter space, thus requiring the existence of strictly positive constant  $\rho$  and addition of non-negative definite matrix  $V_0$  in defining the norm  $\|f_{\mathbf{b}}\|_{n, \beta^*} = n^{-1} \sum_{i=1}^n \int_0^T (\mathbf{b} - \beta^*) \{ \mathbf{V}_n(\beta^*) + \rho V_0 \} (\mathbf{b} - \beta^*)$ .

By taking inherently different approach we were able to identify two type of lower and upper bounding processes  $\mathcal{G}_{n,\eta_1}(\mathbf{b})$ ,  $\mathcal{G}_{n,\eta_2}(\mathbf{b})$ , as defined in (13), where in one bound  $\eta_1 = \exp\{-2a_{\mathbf{b}-\beta^*}\}$  and in the other  $\eta_1 = \underline{\omega}$ . Different from classical LAN approach of Le'Cam, from Propositions 1 and 2, we can see that both  $\eta_1$ 's principally depend on  $p$ , where they become zero as  $p$  increases. For local neighborhood  $\mathbb{B}(r_n)$  of size  $r_n > 0$ , defined as  $\{\mathbf{b} : \sum_{j=1}^p d^{1/\gamma_j^*} \|\mathbf{b} - \beta^*\|_{\gamma_j} \leq r_n/d\}$ , by Proposition 1 we have  $P(\mathcal{G}_{n,e^{-2Cr_n}}(\mathbf{b}) \leq \mathcal{R}_n(\mathbf{b}) \leq \mathcal{G}_{n,e^{2Cr_n}}(\mathbf{b}), \forall \mathbf{b} \in \mathbb{B}(r_n)) = 1$ , or in other terms

$$\|\mathbf{I}_{p*d} - \{\nabla^2 \mathcal{L}_n(\beta^*)\}^{-1/2} \{\nabla^2 \mathcal{L}_n(\mathbf{b})\} \{\nabla^2 \mathcal{L}_n(\beta^*)\}^{-1/2}\|_{\infty} \leq e^{2Cr_n}, \forall \mathbf{b} \in \mathbb{B}(r_n).$$

The above inequality is log likelihood alternative to irrerepresentable condition (Meinshausen and Bühlmann, 2006) defined to  $l_2$  loss functions, where  $\nabla^2 \mathcal{L}_n(\mathbf{b}) = \nabla^2 \mathcal{L}_n(\beta^*)$  for every  $\mathbf{b} \in \mathcal{R}^{pd}$ . This inequality also shows that geometrically quadratic bounds to the Cox model are not independent of the size of the local neighborhood  $r_n$ . In the following, we discuss the relationship between these bounds and the choice of  $r_n$ .

It is clear that, when  $r_n$  does not converge to zero, the usual upper bound in the right hand side of the last inequality (as assumed for linear models with  $l_2$  criterions), cannot hold for any  $\mathbf{b}$  in Cox model. If we are in local small neighborhood where  $r_n \rightarrow 0$ , then there is hope that such upper bound might hold. In Spokoiny (2012), for the case of generalized linear models (see Theorem 5.7 of cited paper) similar conclusion was made. There  $\eta_1 = \sqrt{1-\delta}$ ,  $\eta_2 = \sqrt{1+\delta}$ ,  $\delta > \delta(r_n) > 0$  such that

$$\|\mathbf{I}_{p*d} - \{\nabla^2 \mathcal{L}_n(\beta^*)\}^{-1/2} \{\nabla^2 \mathcal{L}_n(\mathbf{b})\} \{\nabla^2 \mathcal{L}_n(\beta^*)\}^{-1/2}\|_{\infty} \leq \delta(r_n),$$

for all  $\mathbf{b} \in \mathbb{B}(r_n)$ . If  $r_n$  is fixed and bounded away from zero, then  $\eta_1$  and  $\eta_2$  will be too small and too large respectively. It hence requires uniformly small bound on the value of  $\delta(r_n)$ , which on the other hand grows with  $r_n$ . It shows that previous inequality, needed for sandwich bounds, cannot hold for all values of  $\mathbf{b}$ .

Situation is better for correlated linear models (van de Geer, 2011), where  $\widehat{\Sigma}, \Sigma$ , sample and population covariance matrix respectively, indeed satisfy

$$\|\widehat{\Sigma} - \Sigma\|_{\infty} \leq \sqrt{\log p/n},$$

and finite sample oracle prediction properties of simple linear models carry over to correlated linear models. Previous discussions show that, this may not be possible in the Cox model and that the ideas of local non-asymptotic normality, as in Spokoiny (2012), cannot trivially be extended to high dimensional spaces.

For that end, to extend the idea of local non-asymptotic normality to high dimensional problems, we develop new two stage technique. Firstly with the help of sparsity and regularization, we localize the problem to small, dimensionality independent local neighborhood, where we show that our penalized estimator  $\widehat{\beta}$  belongs to  $\mathbb{B}(r_n)$  with exponentially big probability (see Theorem 3 in Section 4 for more details, where we prove that  $r_n$  is proportional to  $s^2 \lambda_n^2 / \zeta^2$ ). Secondly we localize the complexity of the problem further by shrinking  $\nabla^2 \mathcal{L}_n(\mathbf{b})$  in that local parameter space to its sparse alternative  $\nabla^2 \mathcal{L}_n(\beta^*)$ . Then, we show how to inherit the finite sample oracle properties of penalized  $l_2$  criterions and behave "normally" (see Theorem 4 in Section 4 where we show that proposed estimator inherits prediction error of penalized least squares estimator).

### 3. GENERAL ORACLE INEQUALITY

In this section we show how general oracle inequalities of quadratic type (Rigollet, 2012; Lecué and Mandelson, 2012) (from now on denoted with GOI) cannot be achieved for Cox proportional hazards models (2) even for low dimensional problems with convex penalty function. We show that under no conditions

on the correlation structure in the covariates, oracle inequalities only seemingly reach prediction properties of an oracle estimator with the expected slow rate of  $\log p/n$  (see Theorem 1). They become only local properties, as to have non-trivial bounds, they require  $\hat{\beta}$  to be in a local neighborhood of the sparse alternative  $\beta^*$  (see Corollary 2). Interestingly, we show that they cannot extend to hold outside of such local neighborhood. It becomes apparent that in order to obtain dimensionality free finite sample prediction results we need to first show some localization property of the estimator (6) (which we leave for Section 4).

Note that from (11) we have that the following representation holds uniformly over  $\mathbf{b}$

$$\mathcal{R}_n(\hat{\beta}) - \mathcal{R}_n(\mathbf{b}) = \frac{1}{2} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 - \frac{1}{2} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + (\mathbf{b} - \hat{\beta})^T \mathbf{h}_n(\beta^*)$$

for  $\mathbf{b}_{\hat{\beta}} = c\hat{\beta} + (c-1)\beta^*$  and  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\beta^*$  for a particular choice of  $c \in (0, 1)$  and  $\tilde{c} = \tilde{c}(\mathbf{b}) \in (0, 1)$ . Note that this quadratic representation inherently depends on the structure of the log likelihood  $\mathcal{L}_n$ , but as long as one can define an empirical norm such that  $-\mathbf{b}^T \{\nabla^2 \mathcal{L}_n(\mathbf{b}^*)\} \mathbf{b} = \|f_{\mathbf{b}}\|_{n, \mathbf{b}_{\beta^*}}^2$ , it will hold for any log or partial log-likelihood. In that sense the result of Theorem 1 presented in (20) will hold for any log or partial-log likelihood structure. Note that the later do not have to have an i.i.d. structure and in that sense GOI result presented here are much more general and hold over wider class of problems. The probability bound in (20) will differ from one example to the other. From the definition of the penalized estimator as the minimizer of penalized empirical risk in (6) we have  $\mathcal{R}_n(\hat{\beta}) + \lambda_n P(\hat{\beta}) \leq \mathcal{R}_n(\mathbf{b}) + \lambda_n P(\mathbf{b})$ . Combining previous we get

$$(18) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 2(\hat{\beta} - \mathbf{b})^T \mathbf{h}_n(\beta^*) + 2\lambda_n(P(\mathbf{b}) - P(\hat{\beta})),$$

for any  $\mathbf{b}$  and  $\mathbf{b}^*, \mathbf{b}_{\hat{\beta}}$  fixed and defined as before.

The following condition is needed for controlling the size of the score vector  $\mathbf{h}_n(\beta^*)$  which is necessary for obtaining risk properties of the estimator  $\hat{\beta}$ .

**Condition 1.** *Following conditions are satisfied:*

(i) *There exists a continuous function  $s^{(0)}$  defined on  $\beta^* \times [0, \tau]$  such that*

$$\sup_{0 \leq t \leq \tau} |S_n^{(0)}(\beta^*, t) - s^{(0)}(\beta^*, t)| < a, \quad \inf_{0 \leq t \leq \tau} |s^{(0)}(\beta^*, t)| > b > 0,$$

*almost surely, for some positive constants  $a, b > 0$  such that  $b > a$ .*

(ii) *There exists a multivariate continuous function  $s^{(1)}$ , population version of  $S_n^{(1)}$ , defined on  $\beta^* \times [0, \tau]$  such that for each  $1 \leq j \leq p$ ,  $1 \leq k \leq d$ ,  $\sup_{t \in [0, \tau]} |\{s^{(1)}\}_{jk}(\beta^*, t)| \leq b_1$ , for some positive constant  $b_1 > 0$ .*

(ii) *Process  $\mathbf{Y}(t) = (Y_1(t), \dots, Y_n(t))^T$  is left continuous with right hand limits and such that  $P(Y_j(t) = 1, 0 \leq t \leq \tau) > 0$  for  $j = 1, \dots, n$ .*

This set of conditions replaces classical conditions used in the asymptotic analysis of estimation properties of the Cox model, such as those presented in Condition 2 of Bradic et al. (2011). They are in particular relaxed to handle high dimensional settings, since they are required only at  $\beta^*$  and not necessarily uniformly over parameter space, compared to classical assumptions of Fleming and Harrington (2005).

The following Lemma 1 guarantees that the convexity of the penalty function  $P(\mathbf{b})$  is bounded by the growth of the linear part of stochastic approximation of the risk function  $\mathcal{R}_n(\mathbf{b})$ . Furthermore, it enables us to conclude that the penalized estimator  $\hat{\beta}$  will be sandwiched between arg min of two penalized Gaussian type processes of the form  $\mathcal{G}_{n, \eta} + \lambda_n P$  with function  $P$  as in (7).

LEMMA 1. *On the event  $\mathcal{E}_n = \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+), \forall j \in \{1, \dots, p\} \right\}$  we have for convex penalty functions defined in (7), that the following holds*

$$(19) \quad \mathcal{G}_{n,0}(0) = \min_{\mathbf{b} \in \mathcal{R}^{p \times d}} \{ \mathcal{G}_{n,0}(\mathbf{b}) + \lambda_n P(\mathbf{b}) \}.$$

Note that Proposition 1 does not require any type of irrepresentable condition because it is a result on the lower linear (and not quadratic) approximation of the penalized process  $\mathcal{R}_n(\mathbf{b}) + \lambda_n P(\mathbf{b})$ . Therefore, we are able to obtain general oracle inequality (GOI from hereon) without any assumptions on the correlation structure in the data. Similar properties have been established for  $l_2$  type of penalized criterions (Gaïffas and Guillaou, 2012; Lecué and Mandelson, 2012; Rigollet, 2012) where empirical  $\|f_{\mathbf{b}}\|_n$  norm was an essence. In the next section we will provide GOI for nonparametric Cox model (1). The following theorem is the main result of this section. It represents general oracle inequality showing that the penalized estimator  $\hat{\beta}$  reaches the risk properties of an oracle estimator measured through the functional norm (9) introduced in Section 1.

THEOREM 1 (GOI). *Let  $\hat{\beta}$  be defined as in (6) with the convex penalty function  $P(\mathbf{b})$  defined in (7). Then, on the event  $\mathcal{E}_n = \cap_{j=1}^p \mathcal{E}_{n,j}$ ,  $\mathcal{E}_{n,j} = \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right\}$ , there exists  $c \in (0, 1)$  and  $\hat{\beta}^* = c\hat{\beta} + (1-c)\beta^*$  such that*

$$(20) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 \leq \min_{\mathbf{b} \in \mathcal{R}^{p \times d}} \{ \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 4\lambda_n P(\mathbf{b}) \}$$

for  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\beta^*$  and  $\tilde{c} = \tilde{c}(\mathbf{b}) \in (0, 1)$ . Moreover, there exist a positive constant  $c_1 > 0$ , such that the event  $\mathcal{E}_n$  satisfies

$$(21) \quad P(\mathcal{E}_n) \geq 1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho'^2(0+)\}.$$

*Proof of Theorem 1.* Utilizing Lemma 1 we have that on the event  $\mathcal{E}_n = \{ \|\mathbf{h}_n(\beta^*)\|_{\infty} \leq \lambda_n \rho'(0+) \}$ , for  $\Delta = \hat{\beta} - \mathbf{b} \in \mathcal{R}^{p \times d}$  the following holds:

$$\begin{aligned} (\beta^*)^T \mathbf{h}_n(\beta^*) - \mathcal{L}_n(\beta^*) = \mathcal{G}_{n,0}(0) &\leq \mathcal{G}_{n,0}(\Delta) + \lambda_n P(\Delta) = \\ &= -(\Delta - \beta^*)^T \mathbf{h}_n(\beta^*) - \mathcal{L}_n(\beta^*) + \lambda_n P(\Delta), \end{aligned}$$

which is equivalent to  $\Delta^T \mathbf{h}_n(\beta^*) \leq \lambda_n P(\Delta)$ . Together with (18) it leads to

$$\begin{aligned} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 &\leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 2\lambda_n P(\hat{\beta} - \mathbf{b}) + 2\lambda_n (P(\mathbf{b}) - P(\hat{\beta})) \\ &\leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 4\lambda_n P(\mathbf{b}), \end{aligned}$$

for some  $\mathbf{b}_{\hat{\beta}} = c\hat{\beta} + (1-c)\beta^*$ ,  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\beta^*$  and  $c, \tilde{c} = \tilde{c}(\mathbf{b}) \in (0, 1)$ . The last inequality holds due to increasing, convex and symmetry properties of the penalty function  $P$  (which leads to subadditivity). With the last inequality the statement of the general oracle inequality (20) is proved.

We are left to compute the probability of the event  $\mathcal{E}_n = \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+), \forall j \in \{1, \dots, p\} \right\}$ . To that end note that

$$(22) \quad \mathcal{E}_n^c \subseteq \bigcup_{j=1}^p \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\infty} \geq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right\},$$

where  $\|\mathbf{h}_{n,j}(\beta^*)\|_\infty = \max_{1 \leq k \leq d} |\{\mathbf{h}_n\}_{jk}(\beta^*)|$ . First note that we can decompose  $\{\mathbf{h}_n\}_{jk}(\beta^*) = v_{jk} + \nu_{jk}$  as follows

$$(23) \quad n^{-1} \sum_{i=1}^n \int_0^\tau (\{\mathbf{E}_n(\beta^*)\}_{jk} - \{\mathbf{e}(\beta^*)\}_{jk}) dM_i + n^{-1} \sum_{i=1}^n \int_0^\tau (\{\mathbf{e}(\beta^*)\}_{jk} - \Psi_k(X_{ij})) dM_i.$$

With this in mind we have that

$$(24) \quad \mathcal{E}_n^c \subseteq \bigcup_{j=1}^p \left\{ \max_{1 \leq k \leq d} |v_{jk}| \geq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \cup \max_{1 \leq k \leq d} |\nu_{jk}| \geq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right\}.$$

In order to get rates of convergence, we will extend the idea of Theorem 3.1 in [Bradic et al. \(2011\)](#) to Hoeffding's type of inequalities for maxima of averages. This will be possible due to the particular structure of the martingale  $\{\mathbf{h}_n\}_{jk}(\beta^*)$ . To that end first note that by the boundedness property of functions  $\Psi$  we know that  $\Psi_k(X_{ij})$  are bounded random variables. Hence, each  $\nu_{jk}$  is a sum of a sequence of i.i.d subgaussian random variables. However, across  $k$ 's i.e. group elements,  $\nu_{jk}$  are not independent random variables. Hence, we apply the extension of Hoeffding's inequality (see for example Lemma 14.15 in [Bühlmann and van de Geer \(2011\)](#)),

$$P \left( \max_{1 \leq k \leq d} |\nu_{jk}| \geq \|M\|_n \sqrt{2 \left( t^2 + \frac{\log(2d)}{n} \right)} \right) \leq 2 \exp\{-nt^2\},$$

where  $\|M\|_n$  is proportional to  $\sqrt{\frac{1}{n} \sum_{i=1}^n M_i^2(\tau)}$ . In the last statement we have utilized uniform boundedness of B-spline basis. Since  $\bar{M}$  is a bounded martingale, we can conclude that there exists another constant  $c > 0$  such that  $\|M\|_n \leq c$ . Hence, we have

$$(25) \quad P \left( \max_{1 \leq k \leq d} |\nu_{jk}| \geq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right) \leq 4d \exp \left\{ -n \frac{\lambda_n^2 d^{2/\gamma_j^*} \rho'^2(0+)}{4c^2} \right\}.$$

Next, we need to develop a Hoeffding's type inequality for martingale sequences in order to handle  $v_{jk}$ 's. First we start by bounding the jumps and the predictable variation of the martingale  $v_{jk}$ . Note that the jumps are bounded by

$$|\Delta v_{jk}| = \frac{1}{n} |\{\mathbf{E}_n(\beta^*)\}_{jk} - \{\mathbf{e}(\beta^*)\}_{jk}| \leq \frac{1}{n} \sup_{0 \leq t \leq \tau} \|\{\mathbf{E}_n(\beta^*, t)\} - \{\mathbf{e}(\beta^*, t)\}\|_\infty.$$

For the predictable variation process we have

$$\begin{aligned} \langle \Delta v_{jk} \rangle_2 &= \frac{1}{n^2} \int_0^\tau [\{\mathbf{E}_n(\beta^*, t)\}_{jk} - \{\mathbf{e}(\beta^*, t)\}_{jk}]^2 d\langle \bar{M}(t) \rangle \\ &\leq \frac{1}{n} \sup_{0 \leq t \leq \tau} \|\{\mathbf{E}_n(\beta^*, t)\} - \{\mathbf{e}(\beta^*, t)\}\|_\infty^2 \int_0^\tau S_n^{(0)}(\beta^*, t) d\Lambda_0(t) \\ &\leq \frac{c_1}{n} \sup_{0 \leq t \leq \tau} \|\{\mathbf{E}_n(\beta^*, t)\} - \{\mathbf{e}(\beta^*, t)\}\|_\infty^2, \end{aligned}$$

where by utilizing Condition 1 there exists a constant  $c_1 > 0$  such that  $\int_0^\tau S_n^{(0)}(\beta^*, t) d\Lambda_0(t) \leq \int_0^\tau s^{(0)}(\beta^*, t) d\Lambda_0(t) + \sup_{0 \leq t \leq \tau} |S_n^{(0)}(\beta^*, t) - s^{(0)}(\beta^*, t)| \Lambda_0(\tau) \leq c_1$ . The following Lemma is crucial in establishing Gaussian quadratic rates in Bernstein's inequality. Its proof is relegated to the Appendix.



LEMMA 2. *If Condition 1 is satisfied the following bound holds almost surely:*

$$(26) \quad \sup_{0 \leq t \leq \tau} \|\mathbf{E}_n(\boldsymbol{\beta}^*, t) - e(\boldsymbol{\beta}^*, t)\|_\infty \leq 2a_1^2,$$

for  $a_1 = b_1 + aC + \sup_t |s^{(0)}(\boldsymbol{\beta}^*, t)|$ ,  $C = \max_{i,j,k} |\Psi_k(X_{ij})|$ .

From Lemma 2 we have that there exist constants  $c_1, K$  and  $c_2 = 2a_1^2$  and  $K_1$  such that  $|\Delta v_{jk}| \leq c_2/n = K$  and  $\langle \Delta v_{jk} \rangle_t \leq c_1 c_2^2/n = K_1^2$ . To that end we can now apply martingale deviation inequality (van de Geer, 1995) to  $v_{jk}$  to obtain

$$P(|v_{jk}| \geq u_n) \leq 2 \exp\{-nu_n^2/(Ku_n + K_1^2)\}.$$

For the choice of  $u_n = \lambda_n \sqrt{n} d^{1/\gamma_j^*} \rho'(0+)$  note that  $Ku_n = \lambda_n c_2 \rho'(0+) d^{1/\gamma_j^*} / \sqrt{n} \leq \lambda_n c_2 d^{1/\gamma_j^*} / \sqrt{n} \leq c_2 d^{1/\gamma_j^*} / \sqrt{n}$ . Note that  $K_1^2 = c_1 c_2^2/n$ . Since  $c_1$  is an upper bound on  $\int_0^\tau S_n^{(0)}(\boldsymbol{\beta}^*, t) d\Lambda_0(t)$  we can always choose it to be larger than  $\sqrt{n}d$  (which for a typical choice of  $d \sim n^{-1/2}$  becomes a constant) and as such it will satisfy that  $K_1^2 = c_1 c_2^2/n \geq c_2 d^{1/\gamma_j^*} / \sqrt{n} \geq Ku_n$ . With the last statement, using union bound we have

$$P\left(\max_{1 \leq k \leq d} |v_{jk}| \geq u_n\right) \leq 2d \exp\{-nu_n^2/2K_1^2\}.$$

Together with (25) we conclude that there exist another positive constant  $c_1 = \max\{\frac{1}{4c^2}, \frac{1}{2K_1^2}\}$  such that

$$\begin{aligned} & P\left(\sum_{k=1}^d |\{\mathbf{h}_n\}_{jk}(\boldsymbol{\beta}^*)| > \lambda_n d^{1/\gamma_j^*} \rho'(0+)\right) \\ & \leq 6d \max \left\{ \exp \left\{ -n \frac{\lambda_n^2 d^{2/\gamma_j^*} \rho'^2(0+)}{4c^2} \right\}, \exp \left\{ -n \frac{\lambda_n^2 d^{2/\gamma_j^*} \rho'^2(0+)}{2K_1^2} \right\} \right\} \\ & \leq 6d \exp\{-c_1 n \lambda_n^2 d^{2/\gamma_j^*} \rho'^2(0+)\}, \end{aligned}$$

for each  $1 \leq j \leq p$ . This implies by simple union bound that

$$P(\mathcal{E}_n) \geq 1 - 6d \sum_{j=1}^p \exp\{-c_1 n \lambda_n^2 d^{2/\gamma_j^*} \rho'^2(0+)\} \geq 1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho'^2(0+)\}.$$

□

The risk properties of an estimator  $\hat{\boldsymbol{\beta}}$  are stated in terms of empirical functional norms  $\|\cdot\|_{n,\hat{\boldsymbol{\beta}}}$ . Note that parameter  $\mathbf{b}^*$  on the right hand side of (20) is truly different for every  $\mathbf{b}$ , as each constant  $\tilde{c} = \tilde{c}(\mathbf{b})$  is different for each  $\mathbf{b}$ . Empirical result of similar kind appeared independently in the Theorem 6.1 of van de Geer (2011) on structured group lasso for linear regression. Because of the complex log likelihood structure we have to sacrifice simplicity and state the results in terms of two different empirical norms. They become equal to each other in the special case of linear models for which case they become equivalent to the empirical norm  $\|\cdot\|_n$ . Lemler (2012) considered GOI results for the Cox model with known baseline hazard in terms of a newly defined empirical Kullback Leibler Divergence (see Theorem 2 of aforementioned work) for Lasso penalized conditional hazards model. It is not clear how that result relates to  $l_2$  empirical norm (10) and how it is affected by dimensionality  $p$ . Opposing their work we establish that the relationship between the two norms might be significantly impaired by the dimensionality  $p$  and that one has to be careful when stating results in terms of empirical norms as they might become trivial and non-sensical.

In more details, note that the obtained GOI result of Theorem 1 is not an exact GOI in the sense that the left hand side and right hand side have leading coefficient equal to one (Lecué and Mandelson, 2012). Although it might look like they do, due to the structure of the empirical norms  $\|\cdot\|_{n, \mathbf{b}_\beta}$  and  $\|\cdot\|_{n, \mathbf{b}^*}$  and Propositions 2 and 1 this is clearly not the case (see Corollary 2). Due to the complexity of log partial likelihood  $\mathcal{L}_n(\mathbf{b})$  it is impossible to obtain an exact GOI in this case. There is very little hope that GOI similar to linear models can be achieved here. More generally, we conjecture this holds for many complex log-likelihoods whose Hessian matrix depends on parameter structure. To that end, notice that from the discussion on Local Non-Asymptotic Quadraticity of Section 2.1 and Local Non-Asymptotic Normality of Section 2.2 we can conclude the following localized versions of the previous Theorem.

**COROLLARY 2 (Local GOI).** *Using the notation and conditions of Theorem 1 the following holds on the event  $\mathcal{E}_n \cap \{\underline{\omega} > 0\}$ , for  $\epsilon = (N - \underline{\omega})/\underline{\omega}$ ,*

$$(27) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq (1 + \epsilon) \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, \mathbf{b} \in \mathbb{B}(r_n)} \left\{ \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 4 \frac{\lambda_n}{N} P(\mathbf{b}) \right\}.$$

with  $\mathbb{B}(r_n) = \{\mathbf{b} \in \mathcal{R}^{pd} : \sum_{j=1}^p d^{1/\gamma_j^*} \|\mathbf{b}_j - \beta_j^*\|_{\gamma_j} \leq r_n/d\}$ , and  $\mathcal{E}_n$  defined in Theorem 1.

Result of Corollary 2 resembles mentioned GOI results for linear regression models, but requires that  $\underline{\omega}$  is bounded away from zero, a condition that cannot be guaranteed to hold in high dimensions. For example, the larger  $p$  gets, the smaller  $\underline{\omega}$  gets. With sufficiently small  $\underline{\omega}$  (for example of the form  $\underline{\omega} = \delta$ ,  $\delta \rightarrow 0+$ ),  $\epsilon$  explodes. Moreover, this condition is connected to assuming strong convexity of the partial likelihood in the whole  $p \times d$  dimensional space. In recent work Negahban et. al (2012) discuss similar problem of global and local strong convexity in generalized linear models, where they reach the same conclusion that such conditions are far less restrictive in local neighborhood  $\mathbb{B}(r_n)$ . They also address the importance of identifying these local neighborhood in a specific model (see Section 4 on the shape and size of these neighborhood for non-parametric Cox model (2)). In that sense (27) is truly a local GOI result that depends on the size of local neighborhood  $\mathbb{B}(r_n)$ . Lemler (2012) does not discuss similar properties of their estimator. We conjecture that it suffers from the same locality issue.

In this sense GOI results of the least square type, like the one in (27), become localized GOI results for complex models. We believe they cannot trivially be strengthened to hold in full generality without restricting dimensionality of the problem, except for a few special cases where empirical norm (9) is independent of  $\mathbf{b}^*$ . For that end, to be able to infer oracle results of Gaussian type, we first need to make connections of the introduced norm (9) to the norm  $\|\cdot\|_{\beta^*}$  (see Proposition 1). This leads to a two stage argument. In the first stage, the local property of the penalized estimator  $\hat{\beta}$  gives the desired dimensionality free connection between the norms  $\|\cdot\|_n$  and  $\|\cdot\|_{\beta^*}$ . In the second step, we could localize the penalized estimator further and infer finite sample  $\|\cdot\|_n^2$  oracle inequalities that hold for  $p \gg n$ .

#### 4. ORACLE INEQUALITIES OF GAUSSIAN TYPE

In this section we expose the details of the mentioned two-stage argument, and show how  $\hat{\beta}$  inherits finite sample predictive properties similar to penalized least square estimator. Localization of the problem in the first stage is done with the help of sparse oracle inequality (Theorem 2) which provides tail bound on the estimator (Theorem 3). Further localization of the second stage is done with the help of newly developed technique of local non-asymptotic normality where we relate the complexity of the local Cox model to the local linear Gaussian regression model (Theorem 4). As

a by-product we show how the censoring rate  $N/n$  interplays with dimensionality  $p$  and affects its predictive power.

Define the following set of restrictions

$$(28) \quad \mathbb{C}_{\mu,\rho} = \{\mathbf{b} \in \mathbb{R}^{p \times d} : \|\rho(\mathbf{b}_{\mathcal{M}_c^c})\|_1 \leq \mu \|\rho(\mathbf{b}_{\mathcal{M}_*})\|_1\},$$

where  $\|\rho(\mathbf{b})\|_1 = P(\mathbf{b})$  is the penalty function defined in (7) and  $\mathcal{M}_* = \mathcal{M}_*(\beta^*) = \{j \in \{1, \dots, p\} : \|\beta_j^*\|_{\gamma_j} \neq 0\}$ ,  $\text{card}\{\mathcal{M}_*\} \leq s$ . The set  $\mathbb{C}_{\mu,\rho}$  consists of all vectors that have support similar to the sparse vector  $\beta^*$  and changes its shape depending on the penalty function  $\rho$  and the choice of  $\gamma_j$ 's. It represents a generalization of a cone constraint condition that appears in work on Lasso problems with  $l_1$  norm being replaced with FGP function to represent the complex group structure present in the model. To prove sparse oracle inequalities we will use condition in the spirit of Bickel et al. (2009); Meier et al. (2009); Lounici et al. (2011). This condition arises naturally when studying finite sample risk properties of Cox model. Depending on the type of penalty functions, it represents uniform adaptation of Restricted Eigenvalue Condition (Bickel et al., 2009). We refer to van de Geer and Bühlmann (2009) for comparison of different kind of compatibility and restricted eigenvalue conditions and their relationships for linear models. Irrepresentable condition, as defined in Meinshausen and Bühlmann (2006) for linear models, takes more complicated structure for Cox type model (see Bradic et al. (2011) for more details) so we refrain from adopting it for non-parametric hazard models.

**Restricted Eigenvalue Assumption RE**( $\mu, s, \gamma$ ): There exists a positive number  $\zeta = \zeta(s) > 0$  such that

$$(29) \quad \min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0, \delta \in (0,1)} \frac{\|\{-\nabla^2 \mathcal{L}_n(\beta^* + \delta \Delta)\}^{1/2} \Delta\|_2}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}} \geq \zeta,$$

where  $\|\{-\nabla^2 \mathcal{L}_n(\beta^* + \delta \Delta)\}^{1/2} \Delta\|_2^2 = \Delta^T \{-\nabla^2 \mathcal{L}_n(\beta^* + \delta \Delta)\} \Delta$  and  $\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}^2 = \sum_{j \in \mathcal{M}_*} \|\Delta_j\|_{\gamma_j}^2$ . Note that the usual scaling factor of  $\sqrt{n}$  disappears in the definition of the restricted eigenvalue condition because it is included in the definition of the empirical norm  $\|f(\cdot)\|_{n,\cdot}^2$  in the numerator. Moreover, compared to RE condition in Bickel et al. (2009) the denominators differ in that  $l_2$  norm is replaced with  $l_{1,\gamma}$  norm. For every  $\delta_1 \in (0, 1)$ ,  $\Delta^T \{-\nabla^2 \mathcal{L}_n(\beta^* + \delta_1 \Delta)\} \Delta = \|f_{\Delta}\|_{n,\tilde{\mathbf{b}}}^2$  for some particular  $\tilde{\mathbf{b}} = \beta^* + \delta_1 \Delta$ . From decomposition of the risk function  $\mathcal{R}_n$  we have

$$\zeta \leq \min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0, \delta \in (0,1)} \frac{\|f_{\Delta}\|_{n,\beta^* + \delta \Delta}}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}} \leq \min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0} \frac{\|f_{\Delta}\|_{n,\tilde{\mathbf{b}}}}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}}.$$

In comparison to least squares procedures where  $\nabla^2 \mathcal{L}_n(\beta^* + \delta \Delta) = \nabla^2 \mathcal{L}_n(\beta^*) = -\mathbf{X}^T \mathbf{X}$  and restricted eigenvalue conditions are defined on the eigenvalues of  $\mathbf{X}^T \mathbf{X}$ , we need to impose uniform eigenvalue bound as in (29) over all  $\delta \in (0, 1)$ . In more details, following Lemma 1 we have that (29) implies

$$(30) \quad \min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0} \frac{\|f_{\Delta}\|_{n,\tilde{\mathbf{b}}}}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}} \geq \min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0} e^{-a_{\Delta}} \frac{\|f_{\Delta}\|_{n,\beta^*}}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}},$$

where  $a_{\Delta}$  was defined in Proposition 1. Therefore, assuming a point-wise lower bound on the minimum eigenvalue of  $-\nabla^2 \mathcal{L}_n(\beta^*)$ , i.e. assuming  $\min_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0} \{\|f_{\Delta}\|_{n,\beta^*} / \|\Delta_{\mathcal{M}_*}\|_{1,\gamma}\} \geq \zeta$  will not guarantee positive lower bound in (30). Condition (29) can be seen as a rescaling of the minimum eigenvalue problem in classical RE condition needed for the complex likelihood structures.

Determining the class of matrices that satisfy RE( $\mu, s, \gamma$ ) condition is an important open question that needs tremendous novel work, as the random process  $\mathbf{V}_n(\beta^* + \delta \Delta, t)$  doesn't belong to

Gaussian random ensemble for any  $t$  and has complicated correlation structure. In particular, with respect to time it has martingale alike structure and with respect to location i.e to  $\beta^* + \delta\Delta$ , it is a function of the matrix  $\sum_{i=1}^n \sum_{q=1}^n \Psi^T(\mathbf{X}_i)\Psi(\mathbf{X}_q)$ . Using Condition 1 and boundedness of functions  $\Psi$ , matrix  $\int_0^\tau \mathbf{V}_n(0, t)d\bar{N}(t)$  will belong to a random matrix ensemble with sub-gaussian tails, which were studied in Zhou (2009) and Raskutti et al. (2010) respectively. More specifically, we can easily extend their results, to conclude that when  $\delta = -\Delta^T\beta^*$  and the sample size is sufficiently large, there exists a positive  $\zeta_\delta > 0$  such that

$$\min_{\Delta \in \mathbb{C}_{\mu, \rho}, \Delta \neq 0} \frac{\|\{-\nabla^2 \mathcal{L}_n(\beta^* + \delta\Delta)\}^{1/2}\Delta\|_2}{\|\Delta_{\mathcal{M}_*}\|_{1, \gamma}} \geq \zeta_\delta.$$

This is however, not sufficient to indicate the sensitivity of **RE** to change in location and how this restricts the family of matrices that satisfies (29). Although an important question, it lies outside of the scope of the current paper. We are now ready to state the main result of this section.

**THEOREM 2 (SOI).** *Let  $\hat{\beta}$  be defined as in (6) with the convex penalty function  $P(\mathbf{b})$  defined in (7) and let the assumption  $RE(7, s, \gamma)$  hold with  $\zeta = \zeta(s)$ . Then, with probability no less than  $1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho^{*2}(0+)\}$ , for some positive constant  $c_1 > 4$ , under Condition 1, there exists a  $c \in (0, 1)$  and  $\hat{\beta}^* = c\hat{\beta} + (1 - c)\beta^*$  such that*

$$(31) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 \leq \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, |\mathcal{M}_*| \leq s} \left\{ 2\|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + \frac{72}{\zeta^2} \lambda_n^2 \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} \right\},$$

with  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1 - \tilde{c})\beta^*$ ,  $\tilde{c} = \tilde{c}(\mathbf{b}) \in (0, 1)$  and where  $|\mathcal{M}_*| = \text{card}\{\mathcal{M}_*\}$ .

We note that, the previous result does not necessarily imply oracle-type properties of prediction error measured through  $l_2$  norm  $\|f_{\hat{\beta}} - f_{\beta^*}\|_n^2$ . If the dimensionality is small enough to guarantee  $\underline{\omega} > 0$  ( $\underline{\omega}$  defined in Section 2.2), then Gaussian like SOI result would hold without additional work. The argument does not hold if  $p \geq n$  for which we develop two-stage technique. We believe similar reasonings will extend to all settings where the complexity of the model is so severe that the local neighborhood  $\mathbb{B}(r_n)$  of Section 2.2 depends on the dimensionality.

**COROLLARY 3 (Local SOI).** *Let notation and conditions of Theorem 2 hold. Then, on the event  $\mathcal{E}_n \cap \{\underline{\omega} > 0\}$  for  $0 < \varepsilon = (N - \underline{\omega})/\underline{\omega}$*

$$(32) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq (1 + \varepsilon) \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, |\mathcal{M}_*| \leq s, \mathbf{b} \in \mathbb{B}(r_n)} \{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + \Delta_n\}$$

with  $\Delta_n = 72\lambda_n^2 \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} / (N\zeta^2)$ .

Note that  $\varepsilon$  is quite large for large  $n$  and  $p$ , making (32) trivial (with similar arguments as Corollary 2). Local quadraticity and local normality of Section 2, give us guiding principles to inherit local gaussian like properties, while at the same time a way to extend them to the whole  $p$  dimensional space. If we are able to bound  $a_{\hat{\beta}-\beta^*}$  from Proposition 1, then we can see that nesting two neighborhoods, one that carries quadraticity and one that carries normality, local properties might extend for the whole space. Bounding  $a_{\hat{\beta}-\beta^*}$  will show that we almost never end up outside of intersection of those two neighborhoods.

To that end we prove the following Theorem 3. It shows that with high probability, the estimator  $\hat{\beta}$  lies within small elliptical neighborhood that carries quadraticity. For appropriate choice of tuning parameter  $\lambda_n$ , area of this elliptical neighborhood becomes dimensionality independent. In that

sense, it provides tight,  $|\mathcal{M}_*|$  dependent bounds in the first stage of introduced two-stage argument (see the end of Section 2).

**THEOREM 3.** *Let Condition 1 and the assumption  $RE(3, s, \gamma)$  hold with  $\zeta = \zeta(s)$ . Then with probability no less than  $1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho^2(0+)\}$ , for some positive constant  $c_1 > 4$ , for  $\hat{\beta}$  defined in (6) we have*

$$(33) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \beta^*}^2 \leq 16 \frac{\lambda_n^2}{\zeta^2} \exp \left\{ 26C \frac{\lambda_n \sqrt{n}}{\sqrt{N} \zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} \right\} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}.$$

*Proof of Theorem 3.* Following the same steps as in the proof of Theorem 2, but instead of considering all  $\mathbf{b}$ , if we concentrate for the special case of  $\mathbf{b} = \beta^*$  we have from (55) for  $\Delta = \hat{\beta} - \beta^*$ ,

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq 4\lambda_n \sum_{j \in \mathcal{M}_*(\mathbf{b})} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}).$$

This implies that  $\sum_{j \in \mathcal{M}_c^*(\mathbf{b})} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) < 3 \sum_{j \in \mathcal{M}_*(\mathbf{b})} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j})$  or that  $\Delta \in \mathbb{C}_3$  as defined in (28). This combined with assumption  $RE(3, s, \gamma)$  in (29) gives

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 \leq 4\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\hat{\beta}_j - \beta_j^*\|_{\gamma_j}) \leq 4\lambda_n \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*} / \zeta,$$

where in the last step we utilized a version of (56) for  $\mathbf{b} = \beta^*$ . Solving previous inequality we have that

$$(34) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 \leq 16 \frac{\lambda_n^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}.$$

Note that from  $RE(3, s, \gamma)$  and previous inequality,

$$\begin{aligned} 16 \frac{\lambda_n^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} &\geq \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 = \frac{N (\hat{\beta} - \beta^*)^T \mathbf{V}_n(\mathbf{b}_{\hat{\beta}}) (\hat{\beta} - \beta^*)}{\|\hat{\beta}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*\|_{1, \gamma}^2} \|\hat{\beta}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*\|_{1, \gamma}^2 \\ &\geq \frac{N}{n} \zeta^2 \|\hat{\beta}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*\|_{1, \gamma}^2, \end{aligned}$$

where we used the notation  $\|\hat{\beta}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*\|_{1, \gamma}^2 = \sum_{j \in \mathcal{M}_*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j}^2$ . From Cauchy Schwartz inequality we have that

$$\sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} \leq \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \sqrt{\sum_{j \in \mathcal{M}_*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j}^2} \leq 4 \frac{\lambda_n \sqrt{n}}{\sqrt{N} \zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}.$$

Knowing that  $\hat{\beta} - \beta^* \in \mathbb{C}_3$  and using the convexity of  $\rho$  we have  $\|\rho(\hat{\beta}_{\mathcal{M}_c^*} - \beta_{\mathcal{M}_c^*}^*)\|_1 \leq 3 \|\rho(\hat{\beta}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*)\|_1$  and hence

$$\begin{aligned} \sum_{j=1}^p d^{1/\gamma_j^*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} &= \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} + \sum_{j \notin \mathcal{M}_*} d^{1/\gamma_j^*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} \\ &\leq 13 \frac{\lambda_n \sqrt{n}}{\sqrt{N} \zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}. \end{aligned}$$

From Proposition 1 and the last inequality we have that  $\|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\beta^*}^2 \leq \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 e^{2a_{\hat{\beta}-\beta^*}}$ , with  $a_{\hat{\beta}-\beta^*} \leq 2 \max_{1 \leq i \leq n} |(\hat{\beta} - \beta^*)^T \Psi(\mathbf{X}_i)|$ . Result of the Theorem follows from noticing

$$\begin{aligned} \max_{1 \leq i \leq n} \sum_{j=1}^p |(\hat{\beta}_j - \beta_j^*)^T \Psi(X_{ij})| &\leq \sum_{j=1}^p \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} \max_{1 \leq i \leq q} \left( \sum_{k=1}^d (\Psi_k(X_{ij}))^{\gamma_j^*} \right)^{1/\gamma_j^*} \\ &\leq C \sum_{j=1}^p d^{1/\gamma_j^*} \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} \leq 13C \frac{\lambda_n \sqrt{n}}{\sqrt{N} \zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}. \end{aligned}$$

□

We are left to show that gaussian type of SOI results extend to hold even with the complex log-likelihood structures. Inside neighborhood of type (34) we can now further localize the problem by using Gaussian processes (from Corollary 1) as sandwich bounds around our log-likelihood process. In that way we show that elliptical neighborhood can be further shrunk into  $l_2$  balls of smaller radius. This procedure constitutes second step in the technique we developed leading to the following main result.

**THEOREM 4 (Global SOI).** *Under assumptions of Theorems 2 and 3, with probability no less than  $1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho^2(0+)\}$ , for some positive constant  $c_1 > 4$ ,*

$$(35) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq (1 + \epsilon)(1 + \delta) \min_{\mathbf{b} \in \mathcal{R}^{p^*d}, |\mathcal{M}_*| \leq s} \{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 3\Delta_n\},$$

for a nonnegative constant  $\epsilon = \frac{N-\omega}{\omega}$ ,  $\delta = \exp\{4Cr_n\} - 1$ ,  $r_n = \frac{\sqrt{n\Delta_n}}{2\zeta} \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}$  and  $\Delta_n = 26 \frac{\lambda_n^2}{N\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$ .

*Proof of Theorem 4.* From Theorem 2 we have

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 \leq 2\|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 + \frac{72}{\zeta^2} \lambda_n^2 \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*},$$

for any vector  $\mathbf{b}$  such that  $s(\mathbf{b}) \leq s$ . Note that any such vector  $\mathbf{b} \in \mathbb{C}_{0,\rho}$ . Moreover, we are only interested in looking at the minimum over local neighborhood, hence to that end notice that

$$(36) \quad \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 \leq \|\mathbf{b}_{\mathcal{M}_*} - \beta_{\mathcal{M}_*}^*\|_{1,\gamma}^2 \max_{\Delta \in \mathbb{C}_{\mu,\rho}, \Delta \neq 0, \Delta \in (0,1)} \frac{\|f_{\Delta}\|_{n,\beta^*+\delta\Delta}^2}{\|\Delta_{\mathcal{M}_*}\|_{1,\gamma}^2} \leq s\zeta_{\max}^2,$$

where  $\zeta_{\max}$  is the upper bound on the maximum eigenvalue problem stated above. Note that due to  $RE(3, s, \gamma)$  assumption, we can see it is of the order of  $\zeta^{-1}$ . By following similar arguments as in Theorem 3 (after equation (34)), we have

$$\sum_{j=1}^p d^{1/\gamma_j^*} \|\mathbf{b}_j - \beta_j^*\|_{\gamma_j} \leq \frac{\zeta_{\max}}{\zeta} \sqrt{\frac{n}{N} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}.$$

Utilizing Proposition 1 and Proposition 2

$$\begin{aligned} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 &\leq e^{2C \frac{\zeta_{\max}}{\zeta} \frac{\sqrt{n}}{\sqrt{N}} \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\beta^*}^2 \\ &\leq N e^{2C \frac{\zeta_{\max}}{\zeta} \frac{\sqrt{n}}{\sqrt{N}} \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}} \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2, \end{aligned}$$

where  $C \geq \max_{k,i,j} |\Psi_k(X_{ij})|$ . On the other hand, from Theorem 3, we have  $a_{\hat{\beta}-\beta^*} \leq 26C \frac{\lambda_n \sqrt{n}}{\sqrt{N}\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$ . Hence, using Proposition 1 and Corollary 1 we are able to conclude

$$\begin{aligned} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 &\geq e^{-26C \frac{\lambda_n \sqrt{n}}{\sqrt{N}\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\beta^*}^2 \\ &\geq \underline{\underline{\omega}} e^{-26C \frac{\lambda_n \sqrt{n}}{\sqrt{N}\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2. \end{aligned}$$

Moreover, combining these last few inequalities we have

$$\begin{aligned} \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 &\leq 2 \frac{N}{\underline{\underline{\omega}}} e^{26C \frac{\lambda_n \sqrt{n}}{\sqrt{N}\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} + C \frac{\zeta_{\max}}{\zeta} \frac{\sqrt{n}}{\sqrt{N}} \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}} \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 \\ &\quad + \frac{72N}{\zeta^2 \underline{\underline{\omega}}} \lambda_n^2 e^{26C \frac{\lambda_n \sqrt{n}}{\sqrt{N}\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \\ &\leq (1+\epsilon)(1+\delta) \{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 3\Delta_n\}, \end{aligned}$$

for  $\epsilon = \frac{N-\underline{\underline{\omega}}}{\underline{\underline{\omega}}}$ ,  $\Delta_n = 26 \frac{\lambda_n^2}{N\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$  and  $\delta - 1 = \exp \left\{ 2C \frac{\sqrt{n}\Delta_n}{\zeta} \sqrt{\sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}} \right\}$ . □

In comparison with similar results obtained for least square problems, previous result differs in presence of the exponential term on the right hand side of (42) and the direct influence of censoring rate  $N/n$ . It leads to inherently different choices of tuning parameter  $\lambda_n$  compared to simple linear regression where the choice is only governed by probability bounds. Here the choice of  $\lambda_n$  leads to a tradeoff between big elliptical neighborhood and small probability bound. Nevertheless, for typical choices of tuning parameter  $\lambda_n$  (see Section 6), the exponential term  $\delta$  becomes bounded by a constant when the censoring is not severe, making this result comparable to the one obtained for least squares setups (see for example Meier et al. (2009); Massart and Meynet (2011); Lounici et al. (2011)).

To that end, note that  $\underline{\underline{\omega}}$  (as defined in Section 1) is much larger than  $\underline{\omega}$  as it is defined at the true point  $\beta^*$  and independent of the dimensionality of parameter space. For that reason  $\epsilon$  is independent of dimensionality  $p$  allowing us to handle exponentially high dimensional problems. Moreover, it is independent of  $n$  and  $s$  if the following holds

$$(37) \quad P \left( \left| \frac{\max_{i \in \mathcal{R}_N} \omega_i(\beta^*)}{\min_{i \in \mathcal{R}_1 \setminus \mathcal{R}_2} \omega_i(\beta^*)} - 1 \right| > c \right) \leq e^{-u_n},$$

where  $c > 0$  and  $u_n(c)$  is a diverging sequence. It is easy to see that with subgaussian covariates  $\mathbf{X}_i$  and functions  $\{\Psi_k\}_{k=1}^d$ , this inequality holds true as long as we assume that the norm  $\|f_{\beta^*}\|_\infty$  is bounded.

This equation (37) can be viewed as non-asymptotic equivalence of Le Cam's "uniform asymptotic negligibility condition", needed for Le Cam's second lemma. Le Cam's second lemma shows that the estimator has asymptotically normal distribution and our general SOI result of Theorem 4 shows that our estimator inherits non-asymptotic predictive behavior of penalized gaussian estimator. In this way penalized estimator  $\hat{\beta}$  behaves "similarly" to penalized least squares estimator in linear regression models. In more details, we have the following result.



COROLLARY 4. *Under assumptions of Theorem 3, with probability no less than  $1 - 6pd \exp\{-c_1 n \lambda_n^2 d^{2/\max \gamma_j^*} \rho'^2(0+)\}$ , for some positive constant  $c_1 > 4$ ,*

$$(38) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq 3(1 + \epsilon)(1 + \delta) 16 \frac{\lambda_n^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*},$$

for  $\epsilon = (N - \underline{\omega})/\underline{\omega}$  and  $\delta = \exp\left\{8 \frac{\lambda_n \sqrt{n} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}}{\sqrt{N} \zeta^2}\right\} - 1$ . Moreover,

$$(39) \quad \|\hat{\beta} - \beta^*\|_{1, \gamma} = \sum_{j=1}^p \|\hat{\beta}_j - \beta_j^*\|_{\gamma_j} \leq 13 \sqrt{\frac{n}{N}} \frac{\lambda_n^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}.$$

Dependence on censoring rate  $N/n$  in all of the results in this section is explicit and bares interests on its own. The last statement (39) shows that model selection properties of nonparametric Cox, up to a inverse factor of censoring rate, match those of penalized simple linear regression problems. In asymptotic studies of risk properties in Cox models, it never appeared explicitly. With the increase in sample size  $n$ , the number of uncensored observation increases and  $N/n$  asymptotically converges to a constant. Hence, studying non-asymptotic properties of estimators with censored data becomes significantly more important. We believe that almost no previous work has retrieved explicit dependence of risk properties of censored models on censoring rate.

## 5. SPARSE NON-CONVEX GROUP SELECTION WITH CONVEX GUIDE

By using the guidance of convex group selection, in this section we propose a non-convex extension of the proposed FGP family of penalties where the convexity of function  $\rho$  does not have to be restricted. Although good risk properties of Lasso and group Lasso have been well documented, their non-convex counterparts have not been much studied from the perspective of finite sample oracle risk bounds. Bradic et al. (2011) studied variable selection properties of a class of non-convex relaxations of  $L_0$  penalty, called folded concave penalties, firstly introduced in Lv and Fan (2009), but do not mention finite sample oracle risk bounds. Moreover, family of non-convex functions needed to be restricted by upper bound on concavity property of  $\rho$ . If an equivalent condition is made for this setting, then all the previous results (Theorems 2 -3) would hold for such functions  $\rho$  (with changed **RE** condition to the one of this section).

It is of interest to see what are the finite sample risk properties of group non-convex relaxations to structured  $L_0$  penalty and whether we can relax concavity restrictions. In this section we give results in support that they show good and in most cases better risk oracle bounds and that the concavity bound can be lifted. Let us start with the definition of class of concave penalties through the following Non-Convex Assumption.

**Non-Convex Assumption NCA:** Let  $\rho(t; \lambda_n)$  be increasing and concave in  $t \in [0, \infty)$ , even with respect to  $t$  and zero at  $t = 0$ . This assumption is generalization of previous classes of concave penalty functions as defined in Lv and Fan (2009), Zhang and Zhang (2012). For better prediction properties we are willing to sacrifice differentiability of penalty function and continuity of regularized estimator. Apart from defining class of penalty functions, we need to restrict our analysis to a class of design matrices  $\mathbf{X}$ , that satisfy non-convex analog to restricted eigenvalue condition **RE**( $\mu, s, \gamma$ ) of Section 4 as follows.

**Non-convex Restricted Eigenvalue Assumption  $\mathbf{RE}(\mu, s, \rho, \gamma)$ :** There exists a positive number  $\zeta_\rho = \zeta_\rho(s) > 0$  such that

$$(40) \quad \min_{\Delta \in \mathbb{C}_{\mu, \rho}, \Delta \neq 0, \delta \in (0, 1)} \frac{\sqrt{\Delta^T \nabla^2 - \mathcal{L}_n(\beta^* + \delta \Delta) \Delta}}{\rho(\|\Delta_{\mathcal{M}_*}\|_{1, \gamma})} \geq \zeta_\rho,$$

where  $\rho^2(\|\Delta_{\mathcal{M}_*}\|_{1, \gamma}) = \sum_{j \in \mathcal{M}_*} \rho(\|\Delta_j\|_{\gamma_j})^2$  and the restriction set  $\mathbb{C}_{\mu, \rho}$  is as defined in (28). Note that the shape of the restriction set changes with the penalty function  $\rho$ . For those convex penalty functions, it becomes a subset of corresponding ones for non-convex choices of  $\rho$ . Hence, restricted eigenvalue condition  $\mathbf{RE}(s, \mu, \rho, \gamma)$  for non-convex  $\rho$  implies  $\mathbf{RE}(s, \mu, \gamma)$  and  $\zeta_\rho^2 \leq \zeta^2$ .

We propose the following two stage estimation scheme. Let the initial estimator be defined as  $l_{1, \gamma}$  penalized empirical risk minimizer

$$\hat{\beta}_o = \arg \min_{\mathbf{b} \in \mathbb{R}^{p \times d}} \{ \mathcal{R}_n(\mathbf{b}) + \lambda_{n, o} \sum_{j=1}^p d^{1/\gamma_j^*} \cdot \|\mathbf{b}_j\|_{\gamma_j} \}.$$

As it is a convex minimization criterion, Theorem 3 applies and we can conclude that there exists a  $c \in (0, 1)$  and  $\hat{\beta}_o^* = c\hat{\beta}_o + (1 - c)\beta^* = \beta^* + c(\hat{\beta}_o - \beta^*)$  such that

$$\|f_{\hat{\beta}_o} - f_{\beta^*}\|_{n, \hat{\beta}_o^*}^2 \leq 16 \frac{\lambda_{n, o}^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}.$$

For such defined  $\hat{\beta}_o^*$ , let us define a random local elliptical neighborhood  $\mathcal{B}(\hat{\beta}_o)$  centered at  $\beta^*$  as

$$\mathcal{B}(\hat{\beta}_o) = \left\{ \mathbf{b} : \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \hat{\beta}_o^*}^2 \leq 16 \frac{\lambda_{n, o}^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} \right\}.$$

Then, we are able to define a non-convex estimator  $\hat{\beta}_\rho$  as a solution to the following stochastic optimization problem

$$(41) \quad \hat{\beta}_\rho = \arg \min_{\mathbf{b} \in \mathcal{B}(\hat{\beta}_o)} \left\{ \mathcal{R}_n(\mathbf{b}) + \lambda_n \sum_{j=1}^p \rho(\|\mathbf{b}_j\|_{\gamma_j}) \right\}.$$

Note that this minimization problem is nontrivial, i.e.  $P(\hat{\beta}_\rho \neq 0) \geq P(\mathcal{E}_n)$ , for  $\mathcal{E}_n$ , defined in Theorem 1, as on it there exists a feasible point in the constraint set  $\mathcal{B}(\hat{\beta}_o)$ . Moreover, zero is not an absorbing state of this scheme. Once group  $j$  has been identified as zero in the first step, it has positive probability of being not equal to zero in the second step.

In more details, we assume that there exists  $j$  such that  $\hat{\beta}_{o, j} = 0$  and  $\beta_j^* \neq 0$ . Then  $\hat{\beta}_{o, j}^* = (1 - c)\beta_j^* \neq 0$  and  $(\mathbf{b}_j - \beta_j^*)^T \mathbf{V}_n(\hat{\beta}_{o, j}^*)(\mathbf{b}_j - \beta_j^*) \neq 0$  with  $\mathbf{V}_n(\hat{\beta}_{o, j}^*)$  being a  $d \times d$  submatrix of  $\mathbf{V}_n(\hat{\beta}_o^*)$  corresponding to  $j$ -th group. This on the other hand implies that  $P(\hat{\beta}_{\rho, j} \neq 0) > 0$  in comparison to many thresholded variants of Lasso procedure (see for example Meinshausen and Yu (2009); Zhou (2010)) that do not have this property. This property will guard us against admitting false positives (that Lasso traditionally exhibits) while pertaining good prediction properties.

**THEOREM 5 (Non-convex global SOI).** *Let  $\hat{\beta}_\rho$  be as defined in (41) with the non-convex penalty function  $\rho$  satisfying Condition NCA, and let the assumption  $\mathbf{RE}(3, s, \rho, \gamma)$  hold for such choice of*

$\rho$  with  $\zeta_\rho = \zeta_\rho(s)$ . Then, for some positive constant  $c_1 > 4$ , with probability  $1 - 6pd \exp\{-c_1 n \lambda_n^2\}$ , under Condition 1,

$$(42) \quad \|f_{\hat{\beta}_\rho} - f_{\beta^*}\|_n^2 \leq (1 + \epsilon)(1 + \delta) \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, |\mathcal{M}_*| \leq s, \mathbf{b} \in \mathcal{B}(\hat{\beta}_o)} \{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 3\Delta_n\},$$

for a nonnegative constant  $\epsilon = \frac{N - \underline{\omega}}{\underline{\omega}}$ ,  $\delta = \exp\{4Cr_n\} - 1$ ,  $r_n = \frac{\sqrt{ns\Delta_n}}{2\zeta_\rho}$  and  $\Delta_n = 26 \frac{s\lambda_n^2}{N\zeta_\rho^2}$ .

Note that Theorem 5 does not require any global optimality properties of the non-convex minimizer and in that sense it does not require any concavity bound on the function  $\rho$ . Actually for most non-convex choices of  $\rho$  the FGP function (7) is neither convex or concave. In that sense the results of Theorem 5 extend to hold even for, computationally untractable and discontinuous, two step  $l_{0,\gamma}$  group penalty

$$l_{0,\gamma}(\mathbf{b}) = \sum_{j=1}^p 1\{\|\mathbf{b}_j\|_{\gamma_j} \neq 0\}.$$

As a corollary of previous theorem, we can conclude that  $l_{0,\gamma}$  regularized estimator  $\hat{\beta}_{l_0}$  with convex guide (defined through (41) with  $\rho = l_0$ ), with probability no less than  $1 - 6pd \exp\{-cn\lambda_n^2\beta_n^{-2}\}$  satisfies the following oracle inequality

$$\|f_{\hat{\beta}_{l_0}} - f_{\beta^*}\|_n^2 \leq \frac{N}{\underline{\omega}} \exp\left\{12\sqrt{2}C \frac{s\lambda_n\sqrt{n}}{\zeta_{l_0}^2\sqrt{N}}\right\} \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, |\mathcal{M}_*| \leq s, \mathbf{b} \in \mathcal{B}(\hat{\beta}_o)} \left\{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 72 \frac{s\lambda_n^2}{N\zeta_{l_0}^2}\right\}.$$

Here  $\beta_n$  is defined as minimum signal strength  $\beta_n = \min\{\|\beta_j^*\|_{\gamma_j}, j \in \mathcal{M}_*\}$ .

The choice of  $\lambda_n$  is governed by the rate at which  $\rho'(0+)$  converges to zero, if we assume differentiability of  $\rho$ . Hence, in those cases  $\lambda_{n,o} \ll \lambda_n$ . Moreover, if the number of dictionary functions  $d$  is smaller than  $\sqrt{n}$ , class of non-convex penalties  $\rho$  that satisfy  $\mathbf{RE}(s, 7, \rho, \gamma)$  with  $n^{-1/\max \gamma_j^*} \zeta^2 \leq \zeta_\rho^2 \leq \zeta^2$  has smaller upper risk prediction bound. Implication of this result is the following proposition.

**PROPOSITION 3.** Assume that the size of the dictionary functions  $\Psi$ , is smaller than  $\sqrt{n}$ . Let  $\sigma = \sup_{t \in [0, \tau]} \|E\mathbf{V}_n(\beta^*, t)\|_{\max}$  and let  $A > 4$  be a constant. Then, two stage group SCAD penalty, with the choice of  $\lambda_n \geq A\sigma \sqrt{\frac{\log(pd)}{n\rho'(0+)}}$ , will have smaller finite sample oracle risk bound than that of a group Lasso penalty with the choice of  $\lambda_n \geq A\sigma \sqrt{\frac{\log(pd)}{nd}}$ .

On the other hand, convolution of non-convex  $\rho$  and  $l_\infty$  norm, for example, SCAD and  $l_\infty$  norm of Wang et al. (2009), has worse oracle risk properties in the sense of bigger error bound, than that of  $l_1, l_\infty$  of Negahban and Wainwright (2011) in nonparametric Cox models. Further, finite sample risk properties of non-convex penalties can be heuristically discussed as follows.

Since, no single non-convex penalty beats all others, we design a sequence of non-convex group regularized estimators with convex guide. The choice of  $\rho$  is taken as the sequence of following newly proposed non-convex penalties

$$\rho_{\tau_k}(t) = 1\{|t| \geq \tau_k\} + \sqrt{1 - \frac{(|t| - \tau_k)^2}{\tau_k^2}} \cdot 1\{|t| < \tau_k\}.$$

Curvature of an ellipsoid in small neighborhoods achieves selection properties. The sequence of non-convex estimators  $\hat{\beta}_{\rho_{\tau_k}}$  with a sequence of  $\tau_k$ 's chosen as  $\lambda_n, \frac{\lambda_n}{2}, \frac{\lambda_n}{2^2}, \dots, \frac{\lambda_n}{2^k}, \dots$  will lead to

an estimator with prediction properties approximating  $l_0$  penalty. For  $\tau_1 = \lambda_n$  it is equivalent to group lasso estimator and for  $\tau_k = \lambda_n/2^k$  it is a constrained problem as in (41) with constraint set  $\mathcal{B}(\hat{\beta}_{\rho_{\tau_{k-1}}})$ . It can be viewed as the approximation of computationally unachievable but highly desirable  $l_0$  regularized group estimator.

Let us make brief remark on implementation prospective of the proposed estimation scheme. Neighborhood  $\mathcal{B}(\hat{\beta}_o)$ , centered at the true unknown  $\beta^*$ , is not achievable empirically. Hence, we propose to use the same type of elliptical neighborhood now centered at  $\hat{\beta}_o$  with the same radius size, i.e. at

$$\mathcal{B}_{\hat{\beta}_o}(\hat{\beta}_o) = \left\{ \mathbf{b} : \|\mathbf{f}_{\mathbf{b}} - \mathbf{f}_{\hat{\beta}_o}\|_{n, \mathbf{b}_{\hat{\beta}_o}}^2 \leq 16 \frac{\lambda_{n,o}^2}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} \right\}.$$

This will inherently result in a small change in optimization problem (41), i.e. in doubling the size of  $\mathcal{B}(\hat{\beta}_o)$  but will not affect the result of Theorem 5.

The tuning parameter  $\lambda_{n,o}^2$  is chosen by oracle inequality and varies with the choices of  $\gamma_j$ 's (see Section 6). For a typical choice of group Lasso,  $\lambda_{n,o}$  is proportional to  $\sqrt{\log(pd)/nd}$ .

## 6. EXAMPLES

In this section we show particular cases of group penalty functions (7) and show how they relate to previous work. We consider heterogeneous choices of  $\gamma_j$ , for which different choices of  $\lambda_n$  will be appropriate. By allowing hierarchical and structure within each group  $j$  and among groups, we pay the penalty of having to choose larger tuning parameter than that in independent group settings.

**6.1. Hierarchical Selection and CAP.** Previous work applies to a class of more complex additive models where the groups in the additive model may share some but not necessarily all features across groups. That way each function  $f_j$  can be approximated with a different choice of functions  $\Psi$ . In more details, based on prior information, each  $f_j$  can be approximated by  $\mathbf{b}_{\Gamma_j}^T \Psi_{\Gamma_j}$ , where  $\Gamma_j$  is a set of covariates that belong to group  $j$ . Regularized estimator  $\hat{\beta}$  is then defined as the minimizer of

$$\begin{aligned} & -\frac{1}{n} \sum_{i=1}^n \int_0^\tau \sum_{j=1}^G \mathbf{b}_{\Gamma_j}^T \Psi_{\Gamma_j} dN_i(t) + \int_0^\tau \log \left( \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp \left\{ \sum_{j=1}^G \mathbf{b}_{\Gamma_j}^T \Psi_{\Gamma_j} \right\} \right) d\bar{N}(t) \\ & + \sum_{j=1}^G \lambda_{n,j} |\Gamma_j|^{1/\gamma_j^*} \|\mathbf{b}_{\Gamma_j}\|_{\gamma_j}, \end{aligned}$$

where  $|\Gamma_j|$  stands for the cardinality of that set. Note that this setup incorporates classical group Lasso setting as well, where one would select all  $\gamma_j = 2$ . Then, for some constant  $A > 4$  and the choice of

$$\lambda_{n,j} \geq A\sigma \sqrt{\frac{\log(pd)}{n} |\Gamma_j|^{-2/\gamma_j^*}}, \quad \sigma = \sup_{t \in [0, \tau]} \|E\mathbf{V}_n(\beta^*, t)\|_{\max},$$

once  $\mathbf{RE}(3, s, \gamma)$  condition is satisfied, with probability at least of  $1 - 6\{pd\}^{1-A}$ ,

$$(43) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq \frac{N}{\underline{\omega}} \exp \left\{ 12\sqrt{2}C \frac{\sqrt{\sum_{j \in \mathcal{M}_*} \lambda_{n,j} |\Gamma_j|^{1/\gamma_j^*}} \sqrt{\sum_{j \in \mathcal{M}_*} |\Gamma_j|^{1/\gamma_j^*}}}{\zeta^2} \sqrt{\frac{n}{N}} \right\} \times \\ \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, |\mathcal{M}_*| \leq s} \left\{ 2\|\mathbf{f}_{\mathbf{b}} - f_{\beta^*}\|_n^2 + \frac{72}{N\zeta^2} \sum_{j \in \mathcal{M}_*} \lambda_{n,j}^2 |\Gamma_j|^{2/\gamma_j^*} \right\}.$$

Previous oracle inequality is one of the few that discusses high dimensional finite sample properties of CAP family proposed in linear models in [Zhao et al. \(2009\)](#). Proof of this result is a simple modification of the results presented in the paper with  $\lambda_n$  being adaptive to each group  $\Gamma_j$  and is therefore omitted. Block  $l_1/l_\infty$  penalty as proposed in [Negahban and Wainwright \(2011\)](#) is a member of CAP family. They show empirically that if the overlap among groups is not large enough it behaves worse than plain lasso estimator. Since our groups share big part of this structure due to choice of functions  $\Psi$ , it is not surprising that this penalty has the best properties among all members of CAP family.

In comparison to previous results obtained for group Lasso, the choice of  $\lambda_n$  is chosen larger due to the overlapping structure among groups. In nonparametric setting, non-overlapping case of groups (like multi-task learning case in [Lounici et al. \(2011\)](#) for example) cannot hold and expectedly we pay the price in terms of worse prediction properties. Moreover, the implicit assumption on the censoring rate  $N/n$  and the choice of the number of basis functions  $d$ , takes the form of  $s^2 \log(pd) < N$ . The more events we observe, i.e. the smaller the censoring rate is, the bigger number of basis functions we can choose and the larger  $p$  and  $s$  we can handle. In the classical case of  $d \sim n^{-1/2}$  and no censoring  $N = n$ , the previous constraint becomes  $\log(p) \leq \frac{1}{2} \log n + \frac{n}{s^2}$ , which implies that dense problems with  $s \geq n^{1/4}$  can not be efficiently retrieved.

**6.2. Smoothed Selection.** Throughout previous sections we simplified the technical details and left out the smoothing component of the penalty part. Here, we would like to show that the work extends to this situation with a few adaptations needed. For that end, let us define the penalized smoothed estimator as

$$\hat{\beta}_{\mathbb{S}} = \arg \min_{\mathbf{b}} \left\{ \mathcal{R}_n(\mathbf{b}) + \lambda_n \sum_{j=1}^p \sqrt{d} \rho \left( \|\mathbf{b}_j^T \mathbf{R}_j\|_{\gamma_j} + \sqrt{\mathbf{b}_j^T \mathbf{M}_j \mathbf{b}_j} \right) \right\}, \quad \text{for } \gamma_j \geq 2,$$

for convex and subadditive choice of function  $\rho$  and where the smoothing matrix  $\mathbf{M}_j \in \mathcal{R}^{d \times d}$  ([Meier et al., 2009](#)) contains the inner products of the second derivatives of the B-spline basis functions, that is,  $\{M_j\}_{kl} = \int \Psi_k''(x_j) \Psi_l''(x_j) dx_j$ ,  $k, l = 1, \dots, d$  and  $\mathbf{R}_j \in \mathcal{R}^{d \times d}$  is a matrix obtained from Cholesky decomposition of  $\mathbf{M}_j$  i.e.  $\mathbf{M}_j = \mathbf{R}_j^T \mathbf{R}_j$ . Note that the smoothness of functions  $\Psi$ , require  $\gamma_j$ 's to be chosen larger than the order of desired smoothness of functions  $\Psi$ , as too few number of basis functions cannot guarantee smoothness. Then we can rewrite the problem as

$$(44) \quad \hat{\beta}_{\mathbb{S}} = \arg \min_{\tilde{\mathbf{b}}} \left\{ \mathcal{R}_n(\tilde{\mathbf{b}}) + \lambda_n \sum_{j=1}^p \sqrt{d} \rho \left( \|\tilde{\mathbf{b}}_j\|_{\gamma_j} + \|\tilde{\mathbf{b}}_j\|_2 \right) \right\},$$

with  $\tilde{\mathbf{b}}_j = \mathbf{R}_j \mathbf{b}_j$  and  $\mathcal{R}_n(\tilde{\mathbf{b}}) = -\frac{1}{n} \sum_{i=1}^n \int_0^\tau \sum_{j=1}^p \tilde{\mathbf{b}}_j^T \mathbf{R}_j^{-1} \Psi(X_{ij}) dN_i(t) + \int_0^\tau \log \mathcal{S}_n^{(0)}(\tilde{\mathbf{b}}, t) d\bar{N}(t)$ ,  $\mathcal{S}_n^{(0)}(\tilde{\mathbf{b}}, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\{\sum_{j=1}^p \tilde{\mathbf{b}}_j^T \mathbf{R}_j^{-1} \Psi(X_{ij})\}$ . Crucial part of extending previous results to this novel setting requires extending results of Lemma 1 to the new penalty structure, stated in the following Lemma.

LEMMA 3. *On the event  $\mathcal{E}_n = \{\|\tilde{\mathbf{h}}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq 2\lambda_n \sqrt{d}\rho'(0+), \forall j \in \{1, \dots, p\}\}$ , with  $\tilde{\mathbf{h}}_{n,j}(\beta^*) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau (E_{n,j}(\beta^*, t) - \mathbf{R}_j^{-1} \Psi(X_{ij})) dN_i(t)$ , result of Lemma 1 holds for penalty used in (44). Moreover, results of Propositions 1 and 2 hold for this new empirical risk function with  $a_\mathbf{v}$  and  $\underline{\omega}$  respectively substituted with  $a_\mathbf{v} \times \min_{j \in \{1, \dots, p\}} \{\lambda_{\min}(\mathbf{R}_j)\}$  and  $\underline{\omega}_\mathbb{S}$  defined as*

$$\underline{\omega}_\mathbb{S} := \min_{i \in \{1, \dots, n\}, t \in \cup_{q=1}^n \mathcal{R}_q} \left\{ \frac{\sum_{q=1}^N \exp\{\sum_{j=1}^p \mathbf{b}_j^T \mathbf{R}_j^{-1} \Psi(X_{ij})\} 1\{i \in \mathcal{R}_q\}}{\sum_{l \in \mathcal{R}_q} \exp\{\sum_{j=1}^p \mathbf{b}_j^T \mathbf{R}_j^{-1} \Psi(X_{lj})\}} \right\}.$$

With the help of this result, following previous proofs and closely monitoring triangular inequalities and duplication of constants due to bigger penalization, we can conclude that for some constant  $A > 4$  and the choice of

$$\lambda_n \geq A\sigma \sqrt{\frac{\log(pd)}{nd}}, \quad \sigma = \sup_{t \in [0, \tau]} \|E\mathbf{V}_n(\beta^*, t)\|_{\max},$$

once  $\mathbf{RE}(3, s, 2)$  condition is satisfied, with probability at least of  $1 - 6\{pd\}^{1-A}$ ,

$$(45) \quad \|f_{\hat{\beta}_s} - f_{\beta^*}\|_n^2 \leq \frac{N}{\underline{\omega}_\mathbb{S}} \exp\left\{48C \frac{s\lambda_n \sqrt{n}}{\zeta^2 \sqrt{N}}\right\} \min_{\mathbf{b} \in \mathcal{R}^{p^*d}, |\mathcal{M}_*| \leq s} \left\{2\|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + \frac{144}{N\zeta^2} s\lambda_n^2\right\}.$$

The previous result is a unique finite sample result on prediction properties of non-parametric smoothed estimator in high dimensional Cox model. Although tackled as the last problem, its importance lies in inadmissibility of such results with techniques that already exist in the literature.

## APPENDIX A. APPENDIX

### A.1. Proofs of Lemmas.

*Proof of Lemma 1.* This can be seen from the following reasoning. Let us define a function

$$f(\mathbf{b}) := \mathcal{G}_{n,0}(\mathbf{b}) + n\lambda_n P(\mathbf{b}) = -(\mathbf{b} - \beta^*)^T \mathbf{h}_n(\beta^*) + \lambda_n P(\mathbf{b}) - \mathcal{L}_n(\beta^*).$$

First, we will show that zero is a local minimum of the function  $f(\mathbf{b})$  for all  $\mathbf{b} \in \mathcal{B}(0, 1)$  such that  $\|\mathbf{b}_j\|_1 \leq 1$ . Note that  $f(\mathbf{b}) - f(\mathbf{0}) = \sum_{j=1}^p -\mathbf{b}_j^T \mathbf{h}_{n,j}(\beta^*) + \lambda_n d^{1/\gamma_j^*} \rho(\|\mathbf{b}_j\|_{\gamma_j})$  and conditional on the event  $\mathcal{E}_{n,j} = \{\|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+)\}$ ,

$$\begin{aligned} f_j(\mathbf{b}_j) - f_j(\mathbf{0}_j) &= -\mathbf{b}_j^T \mathbf{h}_{n,j}(\beta^*) + \lambda_n d^{1/\gamma_j^*} \rho(\|\mathbf{b}_j\|_{\gamma_j}) \\ &> \|\mathbf{b}_j\|_{\gamma_j} \left( -\|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} + \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right) \geq 0, \end{aligned}$$

where we have utilized Hölder inequality. Hence we can conclude that  $f(\mathbf{b}) - f(\mathbf{0}) > 0$  if the event  $\mathcal{E}_n = \cap_{j=1}^p \mathcal{E}_{n,j}$ . Since  $f$  is a convex function we can conclude that 0 is a global minimum as well. Note that we don't require unicity of minimum.  $\square$

*Proof of Lemma 2.* From Condition 1 and the following decomposition:  $\|\mathbf{E}_n(\boldsymbol{\beta}^*, t) - e(\boldsymbol{\beta}^*, t)\|_\infty =$

$$(46) \quad \max_{1 \leq j \leq p, 1 \leq k \leq d} \left| \frac{\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)}{S_n^{(0)}(\boldsymbol{\beta}^*, t)} - \frac{\{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)}{s^{(0)}(\boldsymbol{\beta}^*, t)} \right|$$

$$(47) \quad \leq \max_{1 \leq j \leq p, 1 \leq k \leq d} \frac{|\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t) - \{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)|}{|s^{(0)}(\boldsymbol{\beta}^*, t)|}$$

$$(48) \quad + \max_{1 \leq j \leq p, 1 \leq k \leq d} |\{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)| \left| \frac{1}{S_n^{(0)}(\boldsymbol{\beta}^*, t)} - \frac{1}{s^{(0)}(\boldsymbol{\beta}^*, t)} \right|$$

$$(49) \quad \leq \frac{a_1}{b} + \frac{ab_1}{b(b-a)}.$$

To see that the last inequality is correct we follow two arguments stated as (i) and (ii):

(i) From Condition 1 we notice that  $\inf_{0 \leq t \leq \tau} |S_n^{(0)}| \geq b - a > 0$  for  $b > a$ . If  $S_n^{(0)}(\boldsymbol{\beta}^*, t) \geq s^{(0)}(\boldsymbol{\beta}^*, t)$  then utilizing Condition 1 and  $s^{(0)}(\boldsymbol{\beta}^*, t) \geq b > 0$  we conclude  $S_n^{(0)}(\boldsymbol{\beta}^*, t) > b > b - a$ , since  $a \geq 0$ . If  $S_n^{(0)}(\boldsymbol{\beta}^*, t) \leq s^{(0)}(\boldsymbol{\beta}^*, t)$  then notice that  $|S_n^{(0)}(\boldsymbol{\beta}^*, t) - s^{(0)}(\boldsymbol{\beta}^*, t)| = s^{(0)}(\boldsymbol{\beta}^*, t) - S_n^{(0)}(\boldsymbol{\beta}^*, t)$ . From Condition 1 we have that in this case  $s^{(0)}(\boldsymbol{\beta}^*, t) - S_n^{(0)}(\boldsymbol{\beta}^*, t) \leq a$  which is equivalent to  $S_n^{(0)}(\boldsymbol{\beta}^*, t) > b - a$ .

(ii) If Condition 1 holds, then there exists a constant  $a_1 > 0$  such that for every  $1 \leq j \leq p$  and  $1 \leq k \leq d$ ,  $|\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t) - \{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)| < a_1$  almost surely. This easily follows from the following inequality, where  $|\Psi_k(X_{ij})| \leq \tilde{c}_1$  (by boundedness of functions  $\Psi$ ) and  $|s^{(0)}(\boldsymbol{\beta}^*, t)| \leq \tilde{c}_2$  (by assumption on  $|S_n^{(0)}(\boldsymbol{\beta}^*, t) - s^{(0)}(\boldsymbol{\beta}^*, t)| < a < \infty$ ) for some positive constants  $\tilde{c}_1$  and  $\tilde{c}_2$

$$(50) \quad |\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t) - \{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)| \leq |\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)| + |\{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)|$$

$$(51) \quad \leq \tilde{c}_1 |S_n^{(0)}(\boldsymbol{\beta}^*, t)| + b_1$$

$$(52) \quad \leq \tilde{c}_1 |S_n^{(0)}(\boldsymbol{\beta}^*, t) - s^{(0)}(\boldsymbol{\beta}^*, t)| + |s^{(0)}(\boldsymbol{\beta}^*, t)| + b_1$$

$$(53) \quad \leq \tilde{c}_1 a + \tilde{c}_2 + b_1$$

Moreover, notice that  $a_1$  can be taken such that  $a_1 > 1$  by construction in Condition 1, hence (49) is smaller than or equal to  $(a_1 + ab_1)/(b - a) \leq 2a_1^2/(b - a)$ .  $\square$

*Proof of Lemma 3.* We need to adapt Lemma 1 with

$$\begin{aligned} f_j(\mathbf{b}_j) - f_j(\mathbf{0}_j) &= -\mathbf{b}_j^T \mathbf{h}_{n,j}(\boldsymbol{\beta}^*) + \lambda_n \sqrt{d} \rho(\|\mathbf{b}_j\|_{\gamma_j} + \|\mathbf{b}_j\|_2) \\ &\geq \|\mathbf{b}_j\|_{\gamma_j} \left( -\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} + \lambda_n \sqrt{d} \rho'(0+) \left( 1 + \frac{\|\mathbf{b}_j\|_2}{\|\mathbf{b}_j\|_{\gamma_j}} \right) \right). \end{aligned}$$

For  $\gamma_j \geq 2$  we know that  $\|\mathbf{b}_j\|_{\gamma_j} \leq \|\mathbf{b}_j\|_2$  leading to the conclusion that

$$f_j(\mathbf{b}_j) - f_j(\mathbf{0}_j) \geq \|\mathbf{b}_j\|_{\gamma_j} \left( -\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} + 2\lambda_n \sqrt{d} \rho'(0+) \right),$$

which leads us to conclude that the result of Lemma 1 hold for this particular penalty as well on a set  $\mathcal{E}_n = \{\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} \leq 2\lambda_n \sqrt{d} \rho'(0+), \forall j \in \{1, \dots, p\}\}$  whose size is easily deducible from Theorem 1.

Proof of equivalent of Proposition 1 easily extends having at mind that equivalent of  $\mathbf{V}_n(\mathbf{b})$  has extra  $\mathbf{R}_j^{-1}$  terms, which will factor into for example  $a_i$  terms as  $(\mathbf{b} - \boldsymbol{\beta}^*)(\mathbf{R}^{-1} \boldsymbol{\Psi}(\mathbf{X}_i) - \mathbf{E}_n(\boldsymbol{\beta}^*, t))$ .



$\mathbf{R}$  is diagonal block matrix as

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \cdots & \mathbf{0} \\ \vdots & & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_p \end{pmatrix}.$$

With this altered  $\mathbf{V}_n(\mathbf{b})$  the proof follows exact same steps (since each  $\mathbf{R}_j$  is invertible  $d \times d$  matrix with bounded eigenvalues) as proof of Proposition 1 and is therefore omitted.

The definition of weights  $\omega_i(\mathbf{b})$  in the proof of Proposition 2 will be changed to address new weighting matrix  $\mathbf{R}_j$ . Once they are redefined the proof of equivalent of Proposition 2 will follow similar steps, hence we omit the details here.  $\square$

### A.2. Proofs of Propositions.

*Proof of Proposition 3.* For all  $\Delta \in \mathbb{C}_{7,\text{SCAD}}$  and  $\rho = \text{SCAD}$ ,

$$\|\Delta_{\mathcal{M}_*}\|_{1,2}^2 \leq \sum_{j \in \mathcal{M}_*} \rho^2(\|\Delta_j\|_2).$$

Analysis of the upper bound will be split into two parts. First, note that

$$\sum_{j \in \mathcal{M}_*} \rho^2(\|\Delta_j\|_2) \leq \sum_{j \in \mathcal{M}_*} \max \left\{ \rho'(0+), \frac{2\{\rho'\}^2(0+)}{\|\Delta_j\|_2} \right\}^2 \|\Delta_j\|_2^2 \leq \Xi_\rho^2(\Delta) \|\Delta_{\mathcal{M}_*}\|_{1,2}^2,$$

where  $\Xi_\rho(\Delta) = \max \left\{ \rho'(0+), \sup_{j \in \mathcal{M}_*} \left\{ \frac{2\{\rho'\}^2(0+)}{\|\Delta_j\|_2} \right\} \right\}$  by Proposition 1 of Zhang and Zhang (2012). For all  $\Delta \in \mathbb{C}_{7,\text{SCAD}}$  the supremum in the definition of  $\Xi$  is reached among those  $j$  for which  $\|\Delta_j\|_2 \leq \lambda_n$ , and for those  $j$  we already know that  $\rho^2(\|\Delta_j\|_2) = \lambda_n^2 \|\Delta_j\|_2^2 \leq \|\Delta_j\|_2^2$ . On the other hand for all  $\Delta \in \mathbb{C}_{7,\text{SCAD}}$  such that  $\inf_{j \in \mathcal{M}_*} \|\Delta_j\|_2 \geq \lambda_n$ , it is easy to see that if  $\lambda_n n^{1/4} \geq 2\rho'(0+)$  we have

$$\sup_{\Delta \in \mathbb{C}_{7,\text{SCAD}}, \inf_{j \in \mathcal{M}_*} \|\Delta_j\|_2 \geq \lambda_n} \Xi_{\text{SCAD}}(\Delta) \leq n^{1/4}.$$

From Theorem 5 we see that the optimal choice of  $\lambda_n$  for obtaining SOI of group SCAD penalty is  $\lambda_n \geq A\sigma \sqrt{\frac{\log(p\sqrt{n})}{n\rho'(0+)}}$ , (where  $A$  and  $\sigma$  are defined in Section 6) for which the previous requirement becomes trivially satisfied as  $\rho'(0+) \leq \sqrt{n}$  and  $\frac{\log(p\sqrt{n})}{\sqrt{n}} \geq \{\rho'\}^2(0+)$ .  $\square$

### A.3. Proofs of Theorems.

*Proof of Theorem 2.* Let  $\Omega_n = \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+)/2, \forall j \in \{1, \dots, p\} \right\}$ . Following similar steps as with bounding  $\mathcal{E}_n$  in Theorem 1 we know that  $P(\Omega_n) \geq 1 - 6pd \exp \left\{ -c_1 n d^{2-2/\gamma} \lambda_n^2 \rho'^2(0+)/4 \right\}$ . On  $\Omega_n$  we have (from (18) and (19)) that

$$(53) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \left( \rho(\|\hat{\beta}_j - \mathbf{b}_j\|_{\gamma_j}) + 2\rho(\|\mathbf{b}\|_{\gamma_j}) - 2\rho(\|\hat{\beta}_j\|_{\gamma_j}) \right),$$

for all  $\mathbf{b}_{\hat{\beta}} = c\hat{\beta} + (1-c)\beta^*$  and  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\beta^*$  as defined in Theorem 1. From hereon they will be fixed. The proof follows lines of work on SOI problems (see for example Bickel et al. (2009);

Lounici et al. (2011)). For all  $\mathbf{b}$  that satisfy  $s(\mathbf{b}) \leq s$ , with  $s(\mathbf{b})$  denoting the sparsity of vector  $\mathbf{b}$ , we have that

$$\begin{aligned}
 (54) \quad & \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\hat{\beta}_j - \mathbf{b}_j\|_{\gamma_j}) \\
 & \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 2\lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \left( \rho(\|\hat{\beta}_j - \mathbf{b}_j\|_{\gamma_j}) + \rho(\|\mathbf{b}\|_{\gamma_j}) - \rho(\|\hat{\beta}_j\|_{\gamma_j}) \right) \\
 & \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 2\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \left( \rho(\|\hat{\beta}_j - \mathbf{b}_j\|_{\gamma_j}) + \rho(\|\mathbf{b}\|_{\gamma_j}) - \rho(\|\hat{\beta}_j\|_{\gamma_j}) \right).
 \end{aligned}$$

From triangular inequality for the FGP function we have  $\rho(\|\mathbf{b}_j\|_{\gamma_j}) \leq \rho(\|\hat{\beta}_j - \mathbf{b}_j\|_{\gamma_j}) + \rho(\|\hat{\beta}_j\|_{\gamma_j})$  leading to

$$(55) \quad \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 4\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}),$$

where  $\Delta = \hat{\beta} - \mathbf{b}$ .

We consider two cases :

- (i)  $4\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \geq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2$ ,
- (ii)  $4\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2$ .

**Case (i)** From (55) we have

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq 8\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}).$$

This implies that  $\sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) < 7 \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j})$  or that  $\Delta \in \mathbb{C}_{7, \rho}$  as defined in (28). For such  $\Delta$ ,  $\|\rho(\Delta_{\mathcal{M}_*})\|_2 = \sqrt{\sum_{j=1}^p \rho^2(\|\Delta_j\|_{\gamma_j})} \leq \|\Delta_{\mathcal{M}_*}\|_{1, \gamma}$  and (29) we have that

$$\begin{aligned}
 (56) \quad & \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq \sqrt{\bar{d}} \sqrt{\sum_{j \in \mathcal{M}_*} \rho^2(\|\Delta_j\|_{\gamma_j})} \leq \sqrt{\bar{d}} \|\Delta_{\mathcal{M}_*}\|_{1, \gamma} \\
 & \leq \frac{\sqrt{\bar{d}}}{\zeta} \min \left\{ \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n, \hat{\beta}^*}, \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n, \mathbf{b}^*} \right\},
 \end{aligned}$$

for  $\bar{d} = \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$ . Hence we obtain

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) \leq \frac{8\sqrt{\bar{d}}}{\zeta} \lambda_n \min \left\{ \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n, \hat{\beta}^*}, \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n, \mathbf{b}^*} \right\},$$

which together with triangular inequality and  $\min\{a + b, c + d\} \leq \min\{a, c\} + d$  for any  $a, b, c, d$  nonnegative entitles us to conclude that

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 \leq \frac{8\sqrt{\bar{d}}}{\zeta} \lambda_n (\omega + \|f_{\beta^*} - f_{\mathbf{b}}\|_{n, \mathbf{b}^*}),$$

for  $\omega = \min\{\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}, \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}^*}\}$ . On the other hand, from (55) (for  $\mathbf{b} = \beta^*$ ) we have  $\omega^2 \leq \|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \hat{\beta}^*}^2 \leq 4\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\hat{\beta}_j - \beta_j^*\|_{\gamma_j})$ . Combined with (56), this gives that

$\omega^2 \leq 4\lambda_n \sqrt{d}\omega/\zeta$  which gives us  $\omega \leq 4\lambda_n \sqrt{d}/\zeta$  and

$$\begin{aligned} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 &\leq \frac{32}{\zeta^2} \bar{d} \lambda_n^2 + \frac{8\sqrt{d}}{\zeta} \lambda_n \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*} \\ (57) \quad &\leq \frac{32}{\zeta^2} \bar{d} \lambda_n^2 + \frac{8\sqrt{d}}{\zeta} \lambda_n \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*} + \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*}^2. \end{aligned}$$

Remember that for case (i) we assumed  $\|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 \leq 4\lambda_n \sum_{j \in \mathcal{M}_*(\mathbf{b})} d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j})$  which with (56) gives us

$$\begin{aligned} \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 &\leq 4 \frac{\sqrt{d} \lambda_n}{\zeta} \min \left\{ \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n,\hat{\beta}^*}, \|f_{\hat{\beta}} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*} \right\} \\ &\leq 4 \frac{\sqrt{d} \lambda_n}{\zeta} (\omega + \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*}) \\ &\leq 16 \bar{d} \frac{\lambda_n^2}{\zeta^2} + 4 \frac{\sqrt{d} \lambda_n}{\zeta} \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*}. \end{aligned}$$

Solving this quadratic inequality gives us that  $\|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*} \leq (2 + \sqrt{5}) \sqrt{s} \lambda_n / \zeta$ . Together with (57) this leads us to:

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 \leq \frac{48 + 8\sqrt{5}}{\zeta^2} \bar{d} \lambda_n^2 + \|f_{\beta^*} - f_{\mathbf{b}}\|_{n,\mathbf{b}^*}^2,$$

which gives us that for every arbitrary  $\mathbf{b}$  we have

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 \leq \frac{72}{\zeta^2} \lambda_n^2 \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} + 2 \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2.$$

**Case (ii)** From (55) we have

$$\begin{aligned} \|f_{\hat{\beta}} - f_{\beta^*}\|_{n,\hat{\beta}^*}^2 + \lambda_n \sum_{j=1}^p d^{1/\gamma_j^*} \rho(\|\Delta_j\|_{\gamma_j}) &\leq 2 \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 \\ &\leq \frac{72}{\zeta^2} \lambda_n^2 \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*} + 2 \|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2. \end{aligned}$$

□

*Proof of Theorem 5.* Probability of event  $\Omega_n$  can be easily derived from Theorem 1 hence we omit the details. From the definition of the non-convex minimizer  $\hat{\beta}_\rho$  we know that for all  $\mathbf{b} \in \mathcal{B}(\hat{\beta}_\rho)$ ,

$$\mathcal{R}_n(\hat{\beta}_\rho) + \lambda_n \sum_{j=1}^p \rho(\|\{\hat{\beta}_\rho\}_j\|_{\gamma_j}) \leq \mathcal{R}_n(\mathbf{b}) + \lambda_n \sum_{j=1}^p \rho(\|\mathbf{b}_j\|_{\gamma_j}),$$

leading to  $\|f_{\hat{\beta}_\rho} - f_{\beta^*}\|_{n,\{\hat{\beta}_\rho\}_{\beta^*}}^2 \leq$

$$\|f_{\mathbf{b}} - f_{\beta^*}\|_{n,\mathbf{b}^*}^2 + 2 \left( \hat{\beta}_\rho - \mathbf{b} \right)^T \mathbf{h}_n(\beta^*) + 2\lambda_n \sum_{j=1}^p \left( \rho(\|\mathbf{b}_j\|_{\gamma_j}) - \rho(\|\{\hat{\beta}_\rho\}_j\|_{\gamma_j}) \right).$$

Now we need to check if equivalence of Lemma 1 holds for non-convex choices of  $\rho$ . Global optimality condition of this lemma certainly won't hold, but the local optimality still holds. As we are interested

in  $\mathbf{b} \in \mathcal{B}(\hat{\beta}_o)$ , local optimality suffices for our needs. Let us now define the event  $\Omega_n$  as  $\Omega_n = \left\{ \|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} \leq \lambda_n/2, \forall j \in \{1, \dots, p\} \right\}$ . On this event, for all  $\mathbf{b} \in \mathcal{B}(\hat{\beta}_o)$  we have

$$-\mathbf{b}_j^T \mathbf{h}_{n,j}(\beta^*) + \lambda_n \rho(\|\mathbf{b}_j\|_{\gamma_j}) > \|\mathbf{b}_j\|_{\gamma_j} \left( -\|\mathbf{h}_{n,j}(\beta^*)\|_{\gamma_j^*} + \lambda_n \right) \geq 0,$$

as all non-convex penalties in this local neighborhood are lower bounded by  $l_1$  penalty. Then, we can conclude that  $\|f_{\hat{\beta}_\rho} - f_{\beta^*}\|_{n, \{\hat{\beta}_\rho\}_{\beta^*}}^2$  is no bigger than

$$\|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + \lambda_n \sum_{j=1}^p \left\| \{\hat{\beta}_\rho\}_j - \mathbf{b}_j \right\|_{\gamma_j} + 2\lambda_n \sum_{j=1}^p \left( \rho(\|\mathbf{b}_j\|_{\gamma_j}) - \rho(\|\{\hat{\beta}_\rho\}_j\|_{\gamma_j}) \right).$$

In local neighborhood  $\mathcal{B}(\hat{\beta}_o)$ , the concave function  $\rho$  is upper bound of absolute value function. Using its concavity and non-increasing property (which imply subadditivity) we have  $\|f_{\hat{\beta}_\rho} - f_{\beta^*}\|_{n, \{\hat{\beta}_\rho\}_{\beta^*}}^2 + \lambda_n \sum_{j=1}^p \rho(\|\{\hat{\beta}_\rho\}_j - \mathbf{b}_j\|_{\gamma_j}) \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 4\lambda_n \sum_{j=1}^s \rho(\|\{\hat{\beta}_\rho\}_j - \mathbf{b}_j\|_{\gamma_j})$ , for some  $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1 - \tilde{c})\beta^*$  and  $\tilde{c} = \tilde{c}(\mathbf{b}) \in (0, 1)$ . Following the same line of reasoning as in the proof of Theorem 2 (following lines after equation (55)), but for concave constraint set  $\mathbb{C}_{7,\rho}$  and concave restricted eigenvalue condition  $\text{RE}(s, 7, \rho, \gamma)$  we have that there exists  $c \in (0, 1)$  and  $\{\hat{\beta}_\rho\}_{\beta^*} = c\hat{\beta}_\rho + (1 - c)\beta^*$  such that

$$(58) \quad \|f_{\hat{\beta}_\rho} - f_{\beta^*}\|_{n, \{\hat{\beta}_\rho\}_{\beta^*}}^2 \leq \min_{\mathbf{b} \in \mathcal{R}^{p \times d}, \mathbf{b} \in \mathcal{B}(\hat{\beta}_o)} \left\{ 2\|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + \frac{72}{\zeta_\rho^2} s \lambda_n^2 \right\}.$$

Gaussian type oracle bounds now easily follow from extension of Theorem 4 to non-convex penalties. This extension can be done with almost no change in the previous techniques on convex parts except that we should be mindful that the penalty norm takes different form.  $\square$

#### A.4. Proofs of Corollaries.

*Proof of Corollary 1.* The upper bounds follows directly from Proposition 2 as it is uniform bound over the whole parameter space. The lower bound follows by repeating the same steps as in Proposition 2 and definition of the weigh vectors  $\omega_i(\beta^*)$  in (15).  $\square$

*Proof of Corollary 2.* From Theorem 1 we have that

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 \leq \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 + 4\lambda_n P(\mathbf{b}),$$

for any vector  $\mathbf{b}$ .

From Proposition 2 we have that for  $\underline{\omega} > c > 0$ ,

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_{n, \mathbf{b}_{\hat{\beta}}}^2 \geq \underline{\omega} \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2, \text{ and } \|f_{\mathbf{b}} - f_{\beta^*}\|_{n, \mathbf{b}^*}^2 \leq N \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2.$$

From the last three inequalities we conclude that  $\underline{\omega} \|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq N \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + 4\lambda_n P(\mathbf{b})$ , that is

$$\|f_{\hat{\beta}} - f_{\beta^*}\|_n^2 \leq \frac{N}{\underline{\omega}} \|f_{\mathbf{b}} - f_{\beta^*}\|_n^2 + \frac{4}{\underline{\omega}} \lambda_n P(\mathbf{b}).$$

If we define  $\epsilon$  to be such that  $1 + \epsilon = N/\underline{\omega}$  then the result in (27) holds.  $\square$

*Proof of Corollary 3.* The proof follows easily from the result of Theorem 2 and similar arguments as in the proof of Corollary 2.  $\square$

*Proof of Corollary 4.* The proof follows from results of Theorem 3 and taking similar steps as in Theorem 4 and is therefore omitted. Second statement is a simple sub result of Theorem 3.  $\square$

## REFERENCES

- Bickel, P, Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics*, **37**, 1705–1732.
- Bradic, J., Fan, J. and Jiang, J. (2011), Regularization for Cox proportional hazards model with NP dimensionality, *The Annals of Statistics*, **36** (9), .....
- Bühlmann, P., and van de Geer, S. (2011), *Statistics for High Dimensional Data*, Springer, p.556
- Bunea, F., Tsybakov, A. and Wegkamp, M. (2009), Sparsity oracle inequalities for the Lasso, *Electronic Journal of Statistics* , ???
- Bunea, F., Tsybakov, A. and Wegkamp, M. (2007), Aggregation for Gaussian Regression, *The Annals of Statistics* , **35**(4),16741697
- Fleming, T. R. and Harrington, D. P. (2005), *Counting Processes and Survival Analysis*, JOHN-WILEY & SONS
- Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of American Statistical Association* **96**(456),1348–1360.
- Gaïffas, S. and Guilloux, A. (2012), High Dimensional Additive Hazards Models and the Lasso, *Electronic Journal of Statistics*, **6**, 522–546.
- Huang, J. (1999), Efficient estimation of the partially additive Cox model, *Annals of Statistics*, **27**(5), 1536–1563
- Jenatton, R., Audibert, J.-Y. and Bach, F. (2011), Structured Variable Selection with Sparsity-Inducing Norms, *Journal of Machine Learning Research*, **12**, 2777–2824
- Kong, S. and Nan, B. (2012), Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. **arXiv:1204.1992**
- Kolar, M. and Lafferty, J. and Wasserman, L. (2011), Union Support Recovery in Multi-task Learning, *Journal of Machine Learning Research*, **12**, 2415–2435
- Le Cam, L. (1960), Locally asymptotically normal families of distributions, *Univ. California Publ. Statist.* **3**, 3798.
- Lecué, G. and Mandelson, S. (2012), General Non-exact Oracle inequalities for classes with sub exponential envelope, *Submitted to the Annals of Statistics*
- Lemler, S. (2012), Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting, **arXiv:1206.5628**
- Lin, D. Y. and Wei, L. J. (1989), The robust inference for the Cox proportional hazards model, *Journal of American Statistical Association*, **84** (408), 1074–1078
- Lui, H. and Zhang, J. (2008), On the  $l_1 - L_q$  regularized regression, *Technical report of Carnegie Mellon*, **arXiv:0802.1517v1**
- Lounici, K., Pontil, M., Tsybakov, A. B. and van de Geer, S. (2011), Oracle inequalities and optimal inference under group sparsity, *The Annals of Statistics* **39**(4), 2164–2204
- Lv, J. and Fan, Y. (2009), A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares, *Ann. Statist.* **37**(6a),3498–3528.
- Massart, P. and Meynet, C. (2011), An  $l_1$ -Oracle Inequality for the Lasso, **arXiv:1007.4791v1**
- Meier, L., van de Geer, S., and Bühlmann L. (2009), High Dimensional Additive Modeling, *The Annals of Statistics* **37**(6b), 3779–3821.
- Meinshausen, N. and Bühlmann, P. (2006), High-dimensional graphs and variable selection with the Lasso, *The Annals of Statistics*, **34**, 1436–1462.

- Meinshausen, N. and Yu, B. (2009), Lasso-type recovery of sparse representations for high-dimensional data, *The Annals of Statistics*, **37**, 246–270.
- Negahban, D. and Wainwright, M. J. (2011), Simultaneous support recovery in high dimensions: benefits and perils of block  $l_1/l_\infty$ - regularization, *IEEE Transactions on Information Theory*, **57**(6), 3841–3863
- Negahban, S. and Ravikumar, P. and Wainwright, M. J. and Yu, B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, **arXiv:1010.2731v1**
- Obozinski, G., Wainwright, M. and Jordan, M. (2010), Support union recovery in high dimensional multivariate regression, *Manuscript* **arXiv:0808.0711**
- Raskutti, G. and Wainwright, M. and Yu, B. (2012), Restricted Eigenvalue Properties for Correlated Gaussian Designs, *Journal of Machine Learning Research*, **11**, 2241–2259
- Raskutti, G. and Wainwright, M. and Yu, B. (2012), Minimax-Optimal Rates for Sparse Additive Models Over Kernel Classes Via Convex Programming, *Journal of Machine Learning Research*, **13**, 389–427
- Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), Sparse Additive Modeling, *Journal of Royal Statistical Society B* **17**(5), 1009–1030.
- Reynaud-Bouret, P. (2006), Compensator and exponential inequalities for some suprema of counting processes, *Statistics and Probability Letters* **76**1514–1521
- Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics* (to appear). **arXiv:0911.2919**.
- Spokoiny, V. (2012), Parametric Estimation. Finite Sample Theory, *ArXiv* **1111.3029v2**
- Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689705.
- Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of Royal Statistical Society: Series B*. **58**(1), 267–288
- VAN DE GEER, S. (1995). Exponential inequalities for Martingales with Application to Maximum Likelihood Estimation for Counting Processes, *Annals of Statistics* **23**(5), 1779–1801.
- VAN DE GEER, S. (2008). High-dimensional generalized linear models and the Lasso, *The Annals of Statistics*, **36**, 614–645.
- VAN DE GEER, S. (2011). The Lasso with within group structure, *to appear at IMS Collections*.
- van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso, *Electronic Journal of Statistics*, **3**, 1360–1392
- Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009), Hierarchically penalized Cox regression with grouped variables, *Biometrika*, **96**(2), 307–322
- Wu, Y. and Li, L. (2010), Asymptotic Properties of Sufficient Dimension Reduction with a Diverging Number of Predictors, *manuscript*.
- Yuan, M. and Lin, L. (2006), Model Selection and Estimation in Regression with Grouped Variables, *Journal of Royal Statistical Society B* **68**(1), 49–67.
- Zhang, H., Liu, H., Wu, Y. and Zhu, J. (2008), Variable selection for multi-category SVM via adaptive sup-norm regularization, *Electronic Journal of Statistics* **2**, 1149–1167.
- Zhang, H-C. and Zhang, T. (2012), A General Theory of Concave Regularization for High Dimensional Sparse Estimation Problems, **1108.4988v2**.
- Zhao, S. D. and Li, Y. (2012), Principled sure independence screening for Cox models with ultra-high-dimensional covariates, *Journal of Multivariate Analysis*, **105**(1), 397–411
- Zhao, P., Rocha, G. and Yu, B. (2009), The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection, *The Annals of Statistics* **37**(6a), 3468–3497.

- Zhou, S. (2009), Restricted eigenvalue conditions on subgaussian random matrices *Technical report, ETH Zurich*
- Zhou, S. (2010), Thresholded Lasso for high dimensional variable selection and statistical estimation, *Technical Report U Michigan Ann Arbor*